Fall 12-18-2015

Classification of Breast Cancer Patients Using Somatic Mutation Profiles and Machine Learning Approaches

Suleyman Vural
University of Nebraska Medical Center

CLASSIFICATION OF BREAST CANCER PATIENTS USING SOMATIC MUTATION PROFILES AND MACHINE LEARNING APPROACHES

by

Suleyman Vural

A DISSERTATION

Presented to the Faculty of
the University of Nebraska Graduate College
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Biomedical Informatics Graduate Program

Under the Supervision of Dr. Chittibabu Guda

University of Nebraska Medical Center
Omaha, Nebraska

December, 2015

Supervisory Committee:
Dr. James Eudy
Dr. San Ming Wang
Dr. Sanjukta Bhowmick

CLASSIFICATION OF BREAST CANCER PATIENTS USING SOMATIC MUTATION PROFILES AND MACHINE LEARNING APPROACHES

Suleyman Vural, Ph.D.

University of Nebraska Medical Center, 2015

Advisor: Chittibabu Guda, Ph.D.

The high degree of heterogeneity observed in breast cancers makes it very difficult to classify cancer patients into distinct clinical subgroups and consequently limits the ability to devise effective therapeutic strategies. In this study, we explore the use of gene mutation profiles to classify, characterize and predict the subgroups of breast cancers.

We analyzed the whole exome sequencing data from 358 ethnically similar breast cancer patients in The Cancer Genome Atlas (TCGA) project. Identified somatic and non-synonymous single nucleotide variants were assigned a quantitative score (*C-score*) that represents the extent of negative impact on the function of the gene. Using these scores with a non-negative matrix factorization method, we clustered the patients into three subgroups. By comparing the clinical stage of patients among the three subgroups, we identified an early-stage-enriched and a late-stage-enriched subgroup. Comparison of the *C-scores* (mutation scores) of these subgroups identified 358 genes that carry significantly higher rates of mutations in the late-stage-enriched subgroup. Functional characterization of these genes revealed important functional gene families that carry a heavy mutational load in the late-state-enriched subgroup. Finally, using the identified subgroups, we also developed a supervised classification model to predict the likely stage of patients, given their mutation profiles, hence provide clinical insights to

help devise an effective treatment plan.

This study demonstrates that gene mutation profiles can be effectively used with machine-learning methods to identify clinically distinguishable subgroups of cancer patients. Genes and gene families that carry a heavy mutational load in late-stage-enriched cancer patients compared to early-stage-enriched subgroup were also identified from functional analysis of genes. The classification model developed in this method could provide a reasonable prediction of the stage of cancer patients solely based on their mutation profiles. This study represents the first use of only somatic mutation profile data to identify and predict breast cancer subgroups and this generic methodology could also be applied to other cancer datasets.

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude and immeasurable appreciation to the following people. Indeed, this work would not be possible without your help and support.

I would like to thank my advisor **Dr. Chittibabu (Babu) Guda** for all the helpful advice, continuous support, and guidance throughout my Ph.D. journey from Albany to Omaha. I am grateful to my supervisory committee, **Dr. James Eudy, Dr. San Ming Wang,** and **Dr. Sanjukta Bhowmick** for their valuable guidance and suggestions. I am indebted to **Dr. Xiaosheng Wang** for his priceless help on my project. I would like to thank my dear friends in the Guda lab at UNMC for all the enjoyable moments and stress relieving discussions. Those were truly invaluable and unforgettable.

I would like to express my gratitude to the Turkish community in Omaha; their help meant a lot to me. I greatly appreciate all the fruitful discussions we had, which taught me priceless life lessons.

My heartfelt thanks to my brother and sister, **Enis** and **Zeyneb Vural**, for their kind wishes and prayers. A very special thanks to my wife **Sevinç Efendi Vural**, for all the endless, unconditional support and encouragement. Also, I would like to thank my son **Mehmet Selim Vural,** your existence is the greatest source of joy in my life. Last but not the least; I would like to extend my deepest thanks to my beloved parents, **Mehmet and Müyesser Vural**, to whom I owe my life. I cannot thank you enough for all the sacrifices you have made for me, and for your endless patience and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

APC	Adenomatous Polyposis Coli
AUC	Area Under the Curve
BC	Breast Cancer
BWA	Burrows-Wheeler Aligner
BWT	Burrows Wheeler Transform
CADD	Combined Annotation Dependent Depletion
CNV	Copy Number Variation
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic acid
DCIS	Ductal Carcinoma In Situ
EM	Expectation Maximization
ER	Estrogen Receptor
EMT	Epithelial-Mesenchymal Transition
FN	False Negative
FP	False Positive
FDR	False Discovery Rate

GERP	Genomic Evolutionary Rate Profiling
GSEA	Gene Set Enrichment Analysis
HER2	Human Epidermal Growth Factor Receptor 2
IHC	Immunohistochemical
IPA	Ingenuity Pathway Analysis
IDC	infiltrating ductal carcinoma
ICGC	International Cancer Genome Consortium
INDEL	Insertion / Deletion
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
LCIS	Lobular Carcinoma In Situ
ML	Machine Learning
MAF	Minor Allele Frequency
miRNA	*micro RNA*
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
NMF	Non-Negative Matrix Factorization

PR	Progesterone Receptors
PCA	Principal Component Analysis
PPV	Positive Predictive Value
RF	Random Forest
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
SIFT	Sorts Intolerant From Tolerant
TCGA	The Cancer Genome Atlas
TNBC	Triple Negative Breast Cancer
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
UV	Ultraviolet
VEP	Variant Effect Predictor
WEKA	Waikato Environment for Knowledge Analysis

WES	Whole exome sequencing
WGS	Whole genome sequencing

Chapter 1

INTRODUCTION

Cancer is the leading cause of death worldwide accounting for 8.2 million deaths in 2012 (International Agency for Research on Cancer, 2014). According to the World Health Organization's latest world cancer statistics, breast cancer leads the cancers by being the most common reason for female mortality (522,000 deaths in 2012) (International Agency for Research on Cancer, 2013). One in four cancers in women is estimated to be a breast cancer. Moreover, since 2008, breast cancer incidents have increased by more than 20% (International Agency for Research on Cancer, 2013).

Breast cancer is a genetically and clinically complex and heterogeneous disease, comprised of multiple factors that are associated with distinctive histological and biological features, clinical presentations and behaviors, and responses to therapy. Hence, the effectiveness of a specific treatment greatly varies among patients. This multifaceted heterogeneity poses a significant classification challenge in the identification of distinct subtypes.

Breast cancer classification is routinely used for tailoring treatment decisions by oncologists, and it is performed according to different schemes based on different criteria. Major classification methods are based on histopathological analysis, grade and stage of tumor, and analysis of gene expression signatures.

With recent advancements in next-generation sequencing (NGS) methods, the current clinical treatment practices for breast cancer classification may benefit from the addition of in-depth understanding of genetic changes in tumors; hence a novel

classification may be achieved. As the genome sequencing costs are getting cheaper using NGS technology, mutational profiles of tumor samples can be compared against those of the normal samples from the same patients to identify somatic mutations that are specific to a particular patient. Such information from hundreds and thousands of patients could be effective used by computational methods to cluster them into clinically distinguishable subgroups and eventually use this information for effective treatment of cancer patients.

In this work we develop a novel breast cancer classification method using machine learning methods and NGS data from whole exome sequencing of several hundred tumor/normal paired breast cancer samples in the TCGA database.

Chapter 2

BACKGROUND

Cancer occurs as a result of the accumulation of point mutations in critical genes, especially those that repair damaged DNA and control cell growth and division, which allows cells to grow and divide uncontrollably to form a tumor. Point mutations may occur spontaneously during DNA replication or caused by mutagens that can be physical, in the form of radiation from UV rays, X-rays or extreme heat, or chemicals such as molecules that can change the base pairs or disrupt the structure of DNA.

There are a number of cellular processes that affect the expression level of genes such as DNA methylation patterns in the genome, histone modifications, transcriptional regulators some of which are transcriptional factors and miRNAs. However, majority of these factors display a consequential effect to mutations in the DNA, while the somatic mutations in a critical set of genes (referred to as driver genes) are considered as the causative factors for cancers. Due to this causative effect, studying the mutational profiles of tumor DNA is particularly attractive in the current era of personal genomics. With the advances made in the field of genome sequencing and computational biology, it is now possible and also cost effective to sequence only about 1.5% of the total human genome corresponding to the protein coding regions (referred to as Exome) and yet get information about 85% of the mutations with large effects on disease-related traits (Choi et al., 2009).

This section will provide the background information about the basic concepts that are used throughout the dissertation. We will explain the next-generation sequencing technology, which produces the raw sequence information that we use in this dissertation;

variant discovery process, followed by explanation about machine learning; and a review of available scoring methods that quantify the mutations' impact on the gene function.

Next-Generation Sequencing (NGS)

Since NGS technology was first discovered, DNA sequencing has brought a total revolution to our current understanding of molecular biology. Since then these technologies have provided tremendous amounts of data, which are full of biological insights waiting to be unraveled. Nucleotide sequencing is the name of the process for determining the order of nucleotides in a given DNA or RNA molecule.

Sequencing has had a rapidly advancing history since its first development by Edward Sanger in 1975. His technique relies on the chain-termination method referred to as Sanger sequencing (Sanger, Nicklen, & Coulson, 1977). Sanger sequencing was established as the first generation sequencing technology and is accepted as the gold standard. The Human Genome Project was completed using this technology with a cost of $3 billion and taking a total of 13 years.

With a demand for cheaper and faster sequencing in the early 2000s, the sequencing field has witnessed rapid improvement with the development of second-generation sequencing, commonly known as next-generation sequencing (NGS) platforms. These platforms can produce high-throughput sequencing, processing millions of DNA or RNA fragments from a sample, thus enabling sequencing of a complete genome in a single day. Modern Sanger sequencing is typically performed by automated capillary sequencing, while the high-throughput NGS technologies are all based on evaluating signals generated during DNA synthesis. Illumina sequencers constitute the

majority of the NGS sequencing market and are the source of the sequencing data used in this dissertation (Zimmerman, 2014), (DeLuca, 2013).

The most notable drawbacks of NGS when compared to earlier sequencing techniques are higher startup costs and the cost of analyzing the generated data, which constitutes a majority of the effort. On the other hand, NGS is not limited to analyzing predefined regions of a genome and is not vulnerable to the inconsistent nature of microarrays.

NGS methods can target an entire genome or only selected regions of a genome, e.g. only coding regions of a genome which is called exome, hence named as either whole genome sequencing (WGS) or whole exome sequencing (WES), respectively. To summarize the overall sequencing process, small fragments of DNA or RNA are sequenced and then either aligned to a reference genome (provided that a reference genome exists) or fed into a *de novo* assembly process to build relatively contiguous regions of the genome for further analysis.

Exome sequencing is the high-throughput sequencing of every exon in the human genome which represents about 1% to 2% of the whole genome (depending on definition of exome) that corresponds to about 180,000 exons from the coding region, yet contains information on 85% of the disease-causing mutations (Choi et al., 2009). When compared to WGS, WES has much lower costs, which makes it possible to perform a standardized experimental procedure for patients suffering from many diseases, and then help to discover the causative factors of the disease. However, this ability is balanced by the challenge of properly aligning the reads to the reference genome or *de novo* assembly,

calling the true variants, annotating the effects of the changes, filtering the false positive calls, identifying plausible variants and experimental validation of results.

Variant Discovery

After the sequence reads are aligned to a reference genome, the next step is variant discovery to identify variable sites in a tumor genome. This procedure can suffer from high error rates that are due to several factors, including base calling step in sequencers, alignment errors or insufficient depth of coverage. Variant discovery, often referred to as variant calling, involves the execution of a number of computational steps in a sequential order. The whole set of computational steps is called a variant calling pipeline, which typically contains an aligner and a variant caller with a number of intermediate data processing steps. As the name suggests, the aligner maps reads to a reference genome and variant caller identifies variant sites and assigns a genotype to the subject(s).

Sequence alignment is a string matching problem and most efficient methods are based on Burrows-Wheeler Transform (BWT), which uses data compression to gain speed and memory efficiency. Among few other aligners, (MOSAIK (Lee et al., 2014), and CUSHAW3 (Liu, Popp, & Schmidt, 2014)) BWT based aligners including Burrows-Wheeler Aligner (BWA) BWA mem (H. Li, 2013), BWA sampe (H. Li & Durbin, 2009) and Bowtie2 (Langmead & Salzberg, 2012) are the most commonly used algorithms. According to a recent publication, making performance comparison of currently in use aligners, BWT based aligners achieve similar results, due to their similar algorithms and out performing other aligners (Cornish & Guda, 2015).

Unlike aligners, there are many variant callers using a variety of algorithms. Please refer to the provided a review of widely used variant callers in the following

section. Main genomic variations that variant callers aim to identify include: single nucleotide variations (SNVs), and small insertion and deletions (INDELs). Based on the information from SNVs and INDELs, the consequent effects on transcription such as splice junctions and splice variants, and on translation such as synonymous and non-synonymous mutations, loss or gain of stop codons, frame shifts, etc., can be identified. Even with well-mapped, aligned and calibrated reads resolving simple single nucleotide variations require sensitive and specific methods and is a challenging task.

SNVs are categorized based on several criteria. Firstly, variations are distinguished by the way they are inherited. If a variation is identified in germline cells i.e. sperm or egg cells, it is called as germline variation, and this variation may be inherited from a parent to an offspring. Alternatively, the genomic variations found in somatic cells are named as somatic mutations. Somatic mutations are not inherited from a parent, rather they are acquired by an individual during his/her life time and not passed on to progeny. Secondly, variations that cause a change in the translated amino acid are called non-synonymous variations, and those that do not change the amino translated acid, are called synonymous variations. Lastly, genomic variations with less than 0.05 minor allele frequency (MAF) are considered rare variants that are associated with diseases and hence are called as mutations.

In this work, we focus on the somatic non-synonymous mutations that occur in the coding region of human genome.

Machine Learning and Data Mining

Machine learning (ML) and data mining are research areas in computer science, where these names often used interchangeably. ML has gained high attention in recent years due to the availability of high-throughput data from biological experiments with parallel advances in the computing power to process and analyze the data using sophisticated computational tools.

Machine learning is defined as a computational method, aimed to build models, identify patterns and other regularities in data, and using the experience received from past information or previous runs available to the learner to improve its performance. The origin of machine learning dates back to 1957, when the perception model invented based on the human brain neurons. On the other hand**,** data mining is a much younger field, first appeared in early 1990s in the database community, which is closely related to and using techniques from machine learning. Data mining aims to extract useful information from a data set and transform it into a desirable structure for further use. Machine learning is widely used in our daily life, example applications include filtering spam emails, weather forecasting and streaming media suggestions based on earlier seen videos, etc.

Some major problems studied in machine learning and data mining that are used in this work include:

Classification: The task of assigning a class for a sample. The number of classes is often small and increasing the number of classes increases the complexity of the task, but it even can be unbound in cases such as text classification or speech recognition applications.

Clustering: Partitioning the samples into homogeneous groups, in such a way that samples in a group are more similar in a specified measure to each other than to those in the other groups.

Feature selection: The process of selecting a subset of relevant features to use in model building.

Machine learning algorithms can be separated in two distinct groups, namely supervised, and unsupervised learning methods.

Supervised learning algorithms use labeled data (where each instance of the data has a known class) to build models and use these models to predict labels of new unseen data points. In case of continuous labels (such as temperature value in weather forecasting) the machine learning task is named as regression and for nominal labels (such as prediction of it will rain or not) the task is called as classification. A simple classification example would be the prediction of whether or not it will rain today, using historical values of temperature, humidity and wind speed, which constitute features, and the labels are “rain” or “no rain”. In this work, we used supervised machine learning techniques with an aim to make class prediction for breast cancer patients using mutation profiles. There are many supervised machine learning algorithms available including: decision trees, artificial neural networks, and support vector machines (Wagner, 2014).

Unsupervised machine learning approaches address the problem of discovering structure in unlabeled datasets, i.e., the instances of data lack class labels. This problem is commonly referred as cluster analysis or clustering. As a simple example, unsupervised clustering of genes based on expression levels to identify co-regulated pathways can be given. We used clustering in the first step of this work to discover similar patients in

terms of genomic mutation profiles. Hierarchical clustering and k-means clustering are the most widely used unsupervised machine learning techniques; however in the particular case of this dissertation those methods are not sufficient to overcome sparseness issue. For a detailed explanation of this issue, please refer to the methods section.

Review of Variation Scoring Methods

The main significance of this work, as mentioned earlier, is to solely use somatic mutations for breast cancer classification purpose. To achieve this goal, first we need to convert the text-based somatic mutation data into a meaningful score, which can be used for further computation in machine-learning. Quantifying the effect of genomic mutations by itself is a complicated task and there are several methods available for this task. Here, we present a brief survey of the widely used recent methods. For a detailed review, please refer to Ritchie and Flicek, 2014 (Ritchie & Flicek, 2014).

Modern sequencing methods yield an extensive list of sequence variations, which makes manual investigation infeasible; therefore, we need algorithms that can predict the effect of the discovered genomic variations. These methods can be categorized according to the underlying algorithm strategy. Firstly, there are several methods using annotations based on overlap with and proximity to functional elements. These tools consider annotations of regulatory elements, including regions of open chromatin, regions marked by histone modification and sequences bound by specific transcription factors. Secondly, biologically informed rule-based annotation methods use the knowledge of relatively better understood functions of particular nucleotide sequences and make allele-specific predictions about the effect of variants. There are numerous software available for this

analysis which include the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2010), ANNOVAR (K. Wang, Li, & Hakonarson, 2010), and SnpEff (Cingolani et al., 2012). Thirdly, we can group methods using annotations based on sequence motifs and constraints estimated from multiple sequence alignments. These methods evaluate the variations using their genomic position and employ the fact that if a variation is discovered in the proximity of a frequently appearing motif or evolutionary conserved region, then it is expected to impart a higher impact on protein function. Examples include the widely used the Genomic Evolutionary Rate Profiling (GERP) (Cooper et al., 2005) and Sorts Intolerant From Tolerant (SIFT) (Ng & Henikoff, 2001) algorithms. Finally, integrative approaches, using supervised learning algorithms, employ an alternative approach by attempting to learn informative annotations or combinations of annotations, by comparing known functional variants with variants for which there is no direct evidence of functional consequences. The main idea here is to use a 'training set' of variants that are labeled as 'damaging' or 'benign' to identify features or combination of features, which can be used to discriminate between two classes and make accurate predictions for unseen variants. This approach has been adopted by several tools such as PolyPhen (Adzhubei et al., 2010) and MutationTaster (Schwarz, Rödelsperger, Schuelke, & Seelow, 2010).

In this work, we used a more recent method named as Combined Annotation Dependent Depletion (CADD) (Kircher et al., 2014), which incorporates both genic and regulatory annotations, as described in the last category above. In contrast to other tools in its category, CADD uses a training set of variants that have become fixed in the human lineage and therefore presumably represent tolerable variations and deleterious variants

that are not observed in human populations. Hence, CADD uses a much larger training set and avoids sampling (ascertainment) biases associated with existing databases of known disease implicated variants.

Chapter 3

LITERATURE REVIEW

Breast Cancer Staging

The most basic and widely used approach in evaluating treatment options for breast cancer patients involves determining the clinical stage. The stage of breast cancer explains the extent of the cancer in the body and determined based on mainly three measures including tumor size, lymph node status and incidence of metastatic growth. Stage I tumors are measured up to two centimeters and no lymph node involvement have been observed. These tumors often called *in-situ* carcinomas and regarded as a better prognosis group. Stage II tumors are considered to be between two to five centimeters, or have spread to the lymph nodes under the arm on the same side as the initiating breast. Like Stage I tumors, Stage II tumors are also generally effectively treatable. Stage III tumors are more than two inches in diameter and lymph nodes are heavily involved, or cancer has spread to other lymph nodes or tissues near the initiating breast. And lastly, Stage IV tumors are noted with a metastasis that has been identified on underarm, internal mammary lymph nodes or other organs of the body ("Breast cancer stages", retrieved from http://www.cancercenter.com/breast-cancer/stages/,", "Breast Cancer Stages", retrieved from http://www.nationalbreastcancer.org/breast-cancer-stages,", "Breast Cancer Staging and Stages", retrieved from http://ww5.komen.org/BreastCancer/StagingofBreastCancer.html,").

Current Breast Cancer Classification

The classification of breast cancer currently involves the evaluation of histological criteria based on morphology and immunohistochemical (IHC) analyses. The traditional parameters such as histological type, tumor size, histological grade and axillary lymph-node involvement have been shown to correlate with clinical outcome and provide the basis for prognostic evaluation (Elston, Ellis, & Pinder, 1999). In addition, IHC markers such as the expression of hormone receptors (estrogen (ER) and progesterone receptors (PR)) and the overexpression and/or amplification of the human epidermal growth factor receptor 2 (HER2) provide additional therapeutic predictive value and have important role in the treatment decision (Harris et al., 2007). As a more modern approach, microarray-based expression analysis of a select gene panel provides a comprehensive molecular taxonomic classification to breast cancer tumors. This approach has emerged as a standard to identify clinically distinguishable molecular subtypes for use in the current clinical practice.

Histopathological Classification

From histopathological perspective, breast cancer can be broadly categorized into *in situ* carcinoma and invasive (infiltrating) carcinoma. Breast carcinoma *in situ* is further categorized into two types as ductal carcinoma *in situ* (DCIS) and lobular carcinoma *in situ* (LCIS), based on growth patterns and cytological features. DCIS is seen significantly more common than LCIS and includes a heterogeneous group of tumors. Moreover, DCIS tumors have traditionally been categorized into five well recognized subtypes based on their architectural features namely Comedo, Cribiform, Micropapillary, Papillary and Solid. On the other hand, invasive carcinoma also includes heterogeneous

group of tumors and is sub-classified into several histological subtypes namely infiltrating ductal, invasive lobular, ductal/lobular, mucinous (colloid), tubular, medullary and papillary carcinomas. From these subtypes the most common is infiltrating ductal carcinoma (IDC), which covers 70-80% of all invasive carcinomas (C. I. Li, Uribe, & Daling, 2005). Further, IDC is also sub-classified according to tumor grade, which is assessed by evaluating the nuclear pleomorphism, glandular/tubule formation and proliferative activity (mitotic index). Three main IDC sub-classes by grade are namely well-differentiated (grade 1), moderately differentiated (grade 2) or poorly differentiated (grade 3) (Lester & Bose, 2009). This classification is done based on three main criteria namely, nuclear pleomorphism, glandular/tubule formation and mitotic rate.

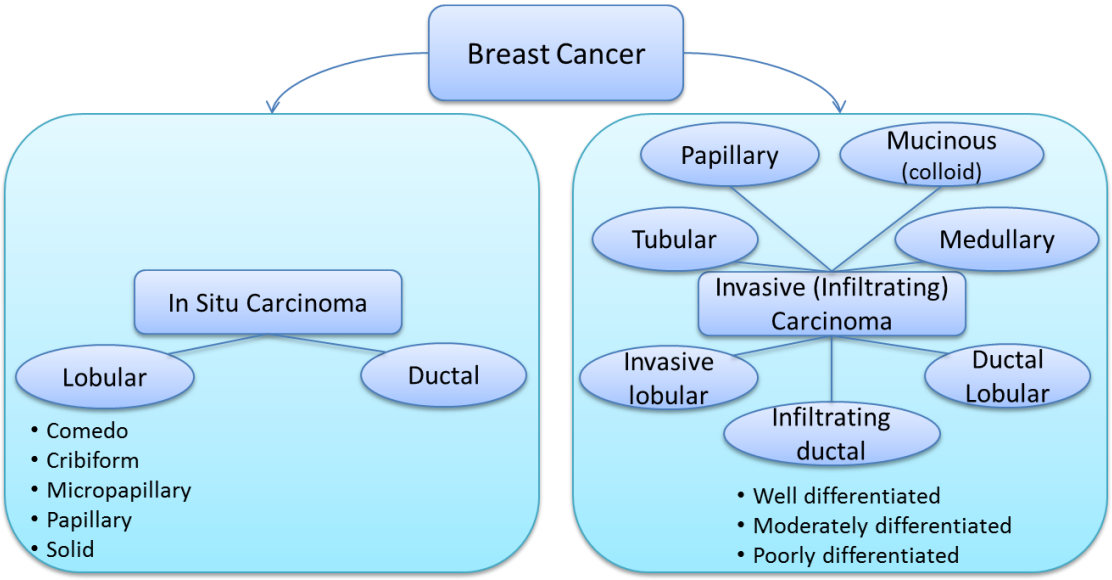


Figure 1: Histological classification of breast cancer subtypes based on architectural features and growth patterns. (Malhotra, Zhao, Band, & Band, 2010)

ER, PR and HER2

In conjunction with histopathological classification, characterization of breast cancers based on the expression of strong biomarkers such as ER, PR, and HER2 has a key role in guiding therapeutic decisions. About 75-80% of all breast cancers are hormone receptor positive, and standardized IHC assays are used to determine the selection of patients for hormone-based therapies. In addition, HER2, an oncogene, is the only predictive marker checked routinely for clinical purpose. Even though there is an inverse association between hormone receptors and HER2, 10-15% of all breast cancers are both hormone receptor and HER2 positive, which are considered to be selected for anti-HER2 based therapies such as the humanized monoclonal HER2 antibody, trastuzumab, which targets the extracellular domain of the HER2 receptor (Konecny et al., 2003). Lastly, the 10-15% of breast cancer patients are recognized by being both hormone receptor (ER&PR) negative and HER2 negative, which are known as triple negative breast cancers (TNBCs). Unfortunately, this type of the breast cancer has the worst prognosis and currently there is not an effective treatment option for TNBCs (Dawson, Provenzano, & Caldas, 2009).

Molecular Classification of Breast Cancers

As a more recent technology, gene expression analysis based on microarray studies gave researchers an opportunity to begin moving towards comprehensive molecular profiling of breast cancer tumors. These studies have led to the discovery of clinically relevant molecular breast cancer subtypes and provided additional insights about the heterogeneity of the disease (Hu et al., 2006; Perou, Sørlie, & Eisen, 2000; Sørlie & Perou, 2001). Application of unbiased hierarchical clustering on gene expression assays has led to the

identification of five distinct breast cancer subtypes (Figure 2) namely Luminal A, Luminal B, HER2 overexpressing, Basal-like and Normal breast tissue-like. Importantly, this molecular classification has successfully discovered sub-classes of ER-positive and/or PR-positive breast cancers as Luminal A and Luminal B. This is a significant achievement because even though clinical assessment of IHC utilizes ER, PR, and HER2 status, these markers could not let the separation of these two distinct subtypes, which have very different clinical outcomes (Sørlie & Perou, 2001; Sørlie & Tibshirani, 2003).

The differences in gene expression patterns in these subtypes reflect the basic alterations in the cell biology of the tumor and are associated with significant variation in clinical outcome such as overall survival and disease free survival (Sørlie & Tibshirani, 2003). Particularly, Luminal A subtype patients are found to have relatively better prognosis while basal-like subtype patients having the worst prognosis.

Following the identification of these intrinsic subtypes, further classification of breast cancers has been proposed. For example, a study conducted on ER-negative tumors has revealed that basal breast cancers are actually a heterogeneous group with at least four main subtypes, and an immune response gene expression module has been discovered, which points to a good prognosis subtype in ER-negative breast cancer (Teschendorff, Miremadi, Pinder, Ellis, & Caldas, 2007). Likewise, a different study has found a new breast cancer intrinsic subtype recognized as Claudin-low or mesenchymal-like (Prat et al., 2010). A characteristic of this subtype is to show an intermediate prognosis between basal and luminal subtypes and to be enriched with cells showing distinct biological properties associated with mammary stem cells and tumour initiating

potential (Bruna et al., 2012; Hennessy & Gonzalez-Angulo, 2009; Lehmann et al., 2011; Lim et al., 2009).

At first, the high cost of gene expression analysis of an abundant number of genes was the obvious obstacle in adoption of the method for clinical purposes. To overcome this, researchers narrowed down the gene list by finding distinct gene signatures for breast cancer subtypes. In one study, investigators have successfully discovered 50 gene signatures, named as PAM50 (Parker et al., 2009; Tibshirani, Hastie, Narasimhan, & Chu, 2002), which can effectively differentiate the molecular subtypes using quantitative real time PCR (qRT-PCR) and is accepted as a replacement for full microarray analysis with the purpose of molecular classification of breast cancers. Moreover, it is demonstrated that using the PAM50 gene set for molecular classification had significantly improved the prediction accuracy for risk of relapse on ER-positive/node negative patients compared to the model that utilizes only clinical variables such as tumor size, axillary lymph-node status and histologic grade.

Besides the identification of intrinsic subtypes, gene expression profiling has also been employed in discovery of distinct prognostic signatures by several groups (Paik, Shak, Tang, & Kim, 2004; Veer, Dai, & Vijver, 2002; Vijver & He, 2002). Mammaprint, which is a microarray-based assay of the Amsterdam 70-gene breast cancer signature, and OncotypeDX, which is a PCR-based assay of a panel of 21 genes, have been approved for clinical use (Cardoso et al., 2008; Sparano & Paik, 2008).

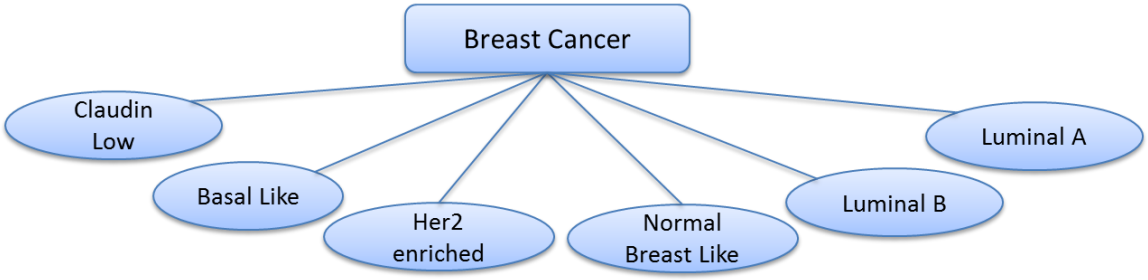


Figure 2: Molecular classification of breast cancer (Malhotra et al., 2010)

Other Hybrid Classification Methods

More recently, thanks to the ease of accessing breast cancer data through projects such as TCGA and organizations like International Cancer Genome Consortium (ICGC), several methods proposing integration of multiple approaches for breast cancer clustering have been published. In 2012, Curtis et al. demonstrated a method combining genome and transcriptome assessments of 2,000 breast cancer patients. By examining the impact of somatic copy number aberrations on the transcriptome, they suggested a novel molecular stratification of breast cancer and revealed novel subgroups (Curtis et al., 2012). Likewise, in 2014, Ali et al. classified breasts cancer into 10 subtypes based on the integration of genomic (copy number variation) and transcriptomic (gene expression) data (Ali et al., 2014). Also in another study, researchers have shown a computational method that combines gene expression and DNA methylation data to implement machine learning aided breast cancer patient classification (List et al., 2014). Lastly, in a recent publication, researchers have proposed a network-based stratification of tumor mutations in which they used somatic mutations as a binary entity in combination with gene

interaction networks and applied non-negative matrix factorization to form four subtypes (Hofree, Shen, Carter, Gross, & Ideker, 2013).

Chapter 4

MATERIALS AND METHODS

Datasets

In development of this project, we downloaded the sequence variation data in variant call format (vcf) for the TCGA breast cancer whole exome sequencing data. Since the data in TCGA comes from a diverse group of patients, to eliminate the population heterogeneity effect, we retrieved a subset of breast cancer patients (n=358) by selecting only white, not Hispanic or Latino patients whose clinical and whole exome sequence data are available. The sequencing data presented in TCGA is processed using several variant callers including, VarScan2 (Koboldt et al., 2012), SomaticSniper (Larson et al., 2012) Samtools (H. Li, Ruan, & Durbin, 2008) . Based on our previous experience with variant callers and supporting literature (Q. Wang et al., 2013), we used the variants discovered by VarScan2. We obtained an average of 17,640 point variations per patient, generated by VarScan2, a highly sensitive tool to detection of somatic mutations in exome sequencing data from normal-tumor pairs.

Data Representation

In this study we used CADD, a method that integrates functional annotations, conservation, and gene-model information into a single score called *C-score*. As mentioned in the original publication, (Kircher et al., 2014) *C-scores* correlate with allelic diversity, annotations of functionality, pathogenicity, disease severity, experimentally measured regulatory effects, and complex trait associations. This score is originally defined to range from negative infinity to positive infinity, where higher score

denotes more deleterious effects; however since our clustering (NMF) algorithm requires all data entries to be positive, we transformed all the scores by adding the minimum score to the original scores.

Our method uses an extensive data structure (mutation score matrix) to keep track of all the deleteriousness scores (*C-scores*) of somatic mutations used for machine learning. The mutation score matrix represents a table that contains the genes in rows and the patients in columns, yielding a matrix of size 18,117 rows by 358 columns, with at least one mutation in each row. And each cell contains the sum of all *C-scores* of mutations found in a gene for a patient. *C-scores* in the mutation score matrix ranges from 0 to 1417.14 and distribution of scores for top 10 variant genes can be seen in Figure 3. (The Python codes developed to build the data structure and apply preprocessing steps are provided in the Appendix.) Comparison against the COSMIC database shows that nine out of these 10 genes (with the exception of FAM38A gene) have evidence of abundant accumulation of somatic mutations in large population screens (Forbes et al., 2014).

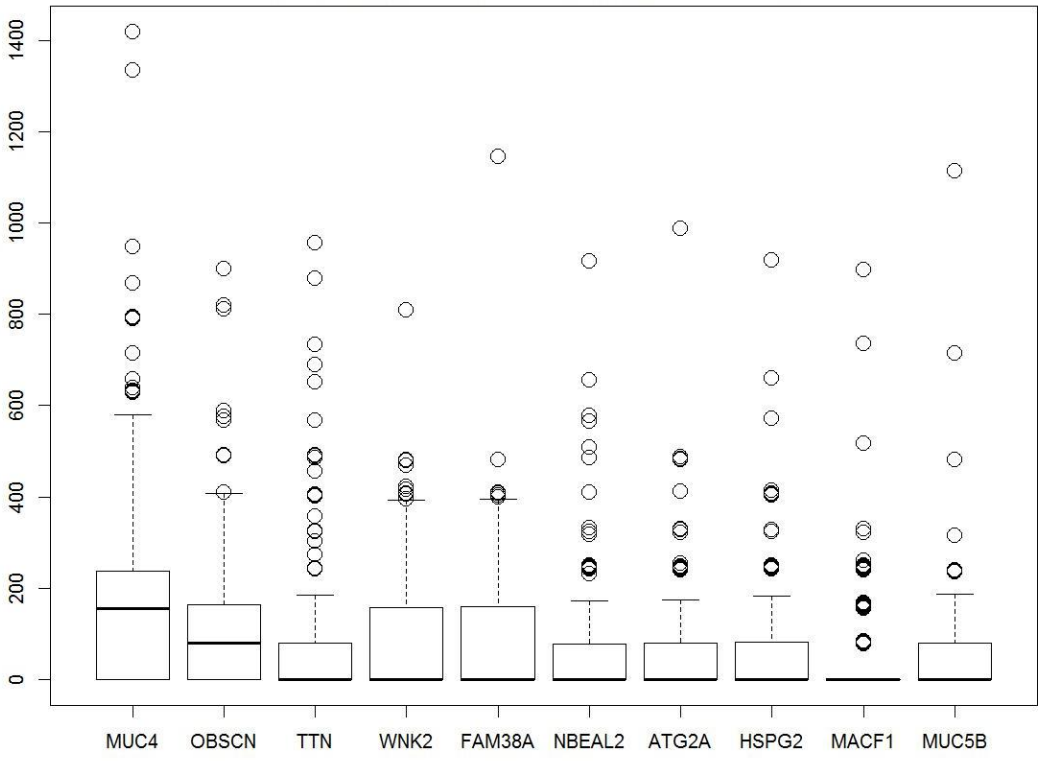


Figure 3: Distribution of total mutational scores for the top 10 variant genes.

Somatic mutation profiles of BC patients exhibit a very sparse data form, unlike other data types such as gene expression or methylation in which nearly all genes or markers are assigned a quantitative value in all the patients**.** Even clinically identical patients may share no more than a single mutation (Bell et al., 2011; Lawrence et al., 2013; TCGA, 2012). Therefore, this problem introduces too many zero valued entries to the main data structure (96%). On the other hand, from machine learning perspective, having a limited number of patients (a far less number of patients than the number of effected genes in the cohort) introduces a dimensionality challenge commonly known as the "curse of dimensionality" in machine learning. Generally machine learning algorithms desire to use a dataset, which has number of cases at least 10 times the

number of features, hence giving a minimum 10:1 sample-to-feature ratio. However, in this study we are faced with a challenge as we observed the sample-to-feature ratio of 1:50 (358/18117) in the main data structure.

In order to overcome the aforementioned challenges, generally there are two popular approaches, namely; feature extraction and feature selection. Feature extraction transforms the current existing features into a lower dimensional space and widely used example methods include principal component analysis (PCA) and linear discriminant analysis (LDA), while feature selection selects a subset of features without applying any transformation. These methods increase the sample-to-feature ratio and decrease the sparseness hence making the clustering both feasible and more effective. In this dissertation we used both feature selection and feature extraction in succession, as further explained in below.

Clustering

We implemented an $m \times n$ mutation score matrix to keep track of the sum of the variant scores in all genes, where *m* is the number of genes (18,117) and *n* is the number of samples (358 patients). The value in entry (*i*, *j*) indicates the mutation score of gene *i* in sample *j*, which is the sum of all *C-scores* of mutations found in the gene *i* for the sample *j*.

We used NMF method for clustering, which aims to find a small number of metagenes, each defined as a positive linear combination of all the genes so that the method can approximate the mutation load of the samples as positive linear combinations of these metagenes. Mathematically, this corresponds to factoring a given non-negative matrix A of size $m \times n$, into two smaller matrices, $W \in \mathbb{R}^{mxk}$ and $H \in \mathbb{R}^{kxn}$, with positive

entries, $A \approx WH$ using a positive integer number $k < min\{m, n\}$. Matrix W, called as a basis matrix and has size $m \times k$, with each of the k columns defining a metagene; and entry w_{ij} represents the coefficient of gene i in metagene j. Matrix H is named as coefficient matrix and has size kxn, with each of the m columns representing the metagene expression pattern of the corresponding sample; and entry h_{ij} represents the mutation load of metagene i in sample j. There are multiple solutions to this problem and in this study we adopt a method by Brunet et al. (Brunet, Tamayo, Golub, & Mesirov, 2004) that was shown to perform better. The solution to form factors W and H can be obtained as explained in the following. The method starts by randomly initializing the matrices W and H and iteratively updates W and H to minimize a divergence function. W and H are updated by using the coupled divergence equations shown in Equation 1.

$$W_{ia} \leftarrow W_{ia} \frac{\sum_u H_{au} A_{iu} / (WH)_{iu}}{\sum_v H_{av}}, H_{au} \leftarrow H_{au} \frac{\sum_i W_{ia} A_{iu} / (WH)_{iu}}{\sum_k W_{ka}}$$

Equation 1: Coupled divergence equations to update the W and H matrices

As a result of factorization, we use coefficient matrix H to group our samples into given number (k) of clusters. Algorithm assigns each sample according to the highest scored metagene in patients designated column in matrix H; meaning that sample *j* will be assigned to the cluster i if h_{ij} is the highest entry in column j.

To specify the optimal number of clusters (rank of clustering) and features (genes) to use in clustering, we used consensus matrix and average silhouette width of consensus matrix.

Since the NMF algorithm starts with a random initial class assignment of samples, repeated runs over the same sample set with constant input parameters may not result in the same sample assigned to the same class between the runs; however, if we observe only a little variation in these associations between runs, then we can conclude with confidence that a strong clustering was performed for this set of parameters (number of clusters and features). This idea forms the basis for our clustering performance evaluations.

Clustering Quality Assessment Methods: Consensus Matrix

Consensus matrix is a concept proposed by Brunet et al. (Brunet et al., 2004) providing visual insights about the performance of clustering. The concept can be explained as follows. In each run, sample to class assignments can be represented by a connectivity matrix C of size mxm by entering $c_{ij} = 1$ if samples i and j are assigned to the same cluster and $c_{ij} = 0$ otherwise. Then the consensus matrix, $\bar{C}$, can be calculated by averaging the connectivity matrix C for many clustering runs. (We selected to use 100) The value in $\bar{C}_{ij}$ ranges from 0 to 1 and reflects the probability of samples i and j assigned to the same cluster. In the case of a stable clustering then we expect to see most of the values in $\bar{C}$ to be close to 0 or 1.

Clustering Quality Assessment Methods: Silhouette Score

In addition to the consensus matrix, we used average silhouette width of consensus matrix (silhouette(consensus)), introduced by Rousseeuw (Rousseeuw, 1987), to quantitatively measure the stability of the clustering runs with different parameters. Silhouette concept is defined as follows: for each sample we can define $a(i)$ as the

average dissimilarity/distance of sample i with all other data within its cluster, the value of $a(i)$ will then indicate how well the sample i fits into its assigned cluster by having a smaller value showing better assignment. Then we can define $b(i)$ by the lowest average dissimilarity of sample i to any other cluster that i is not a member. In other words $b(i)$ indicates the average dissimilarity of sample i to its closest neighboring cluster or its next best fit cluster. Then the silhouette score of a sample can be calculated as in Equation 2 below. The value of $s(i)$ can range from -1 to 1, and being close to 1 means that the sample is perfectly clustered. And average of $s(i)$ over all the samples, named as average silhouette width, shows how well the data has been clustered.

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & \text{if } a(i) > b(i) \end{cases} \quad \text{also can be written as} \quad s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Equation 2: Equation shows how the silhouette score of sample can be computed

Lastly, among several implementations of NMF in various programming languages, we selected to use an R implementation of NMF, published by Gaujoux and Seoighe (Gaujoux & Seoighe, 2010), because of its efficient and flexible parallel processing design and ease of applicability to our study. (The R script preparing the data and running NMF algorithm is provided in the Appendix.)

Feature Selection and Optimization of Clustering

As also mentioned earlier feature selection constitutes an essential step in any machine learning algorithms. We separately performed feature selection for supervised (classification) and unsupervised (clustering) sections of the project.

Due to the higher number of features (tens of thousands genes) being much more than the number of samples (hundreds of samples), we first used feature selection to select only the informative features for clustering; thus to reduce the feature size. We ranked the features in decreasing order of their variance values (Equation 3) and selected top *n* features for clustering.

$$S^2 = \frac{\sum(X - \bar{X})^2}{n - 1}$$

Equation 3: Variance formula

To find the most accurate clustering case, we iteratively run the clustering algorithm over a range of biologically reasonable parameters which is form 2 to 5 clusters and for selected top 10 to 1000 variant genes. Since running the algorithm for each number of cluster and for each 1000 genes would be computationally intensive and not necessary, in finding the correct number of genes we firstly run the algorithm for genes that increase 10 each time (10 genes, 20 genes etc.). And in the second step; we run the algorithm with all the genes in the range around the point we received the highest consensus silhouette score. As an real case example; we receive the highest silhouette score for 850 genes with 3 clusters, in the second step we run the algorithm for all the genes from 840 to 860 genes and found that 856 is actually producing the best clustering. For better and easier understanding, we present this algorithm in the pseudo code below in Table. 1. In finding of Later we used the consensus matrix's silhouette score to determine the optimal number of genes and clusters.

a	for number_of_cluster in 2 to 10:
	  for number_of_features in 10 to 1000 incrementing by 10:
	  Data=select_top_features(Raw_data,number_of_features)
	  NMF (Data,number_of_clusters)

b	number_of_cluster=2
	For number_of_features in 620 to 640 incrementing by 1:
	  Data=select_top_features(Raw_data,number_of_features)
	  NMF (Data,number_of_clusters)

Table 1: Pseudo code for iteratively applying all potential values for k and number of features to keep

Characterization of Clusters

To characterize the clusters we discovered, we correlated the samples in the clusters with their clinical features. For simplicity, we defined stage I and II as early stage and stage III and IV as late stage. The Fisher's exact test was used to assess the stage tendency of clusters.

We compared the mutation score of genes between clusters using the Wilcoxon rank-sum test, and adjusted the multiple testing with the false discovery rate (FDR). The FDR was estimated using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). We used the R language and environment (RDevelopment, 2012) to run all the statistical tests. In addition, we performed functional analysis of the differentially mutated genes between the clusters using the Ingenuity Pathway Analysis ("IPA; Ingenuity Systems Inc.; Redwood, CA, USA," n.d.) and the Gene Set Enrichment Analysis (GSEA) tools (Subramanian, Tamayo, Mootha, Mukherjee, & Ebert, 2005).

Development of Supervised Classification Model

For running feature selection, classification model generation using ML algorithms and performance measurements, we used the Waikato Environment for Knowledge Analysis (WEKA) (Hall et al., 2009) framework, which is an open-source, Java-based framework.

For feature selection, we used the Information gain attribute evaluator (Mitchell, 1997), and Ranker algorithms implemented in Weka for evaluation and searching of the features. We used five diverse and most popular ML algorithms; namely RF (Breiman, 2001), Naïve Bayes (Rish, 2001), C4.5 (named as J48 in Weka) (Salzberg, 1994), SVM (Platt, 1998), and KNN (Stevens, Cover, & Hart, 1967) to build classification models. For performance measurements, we used 10-fold cross-validation. In 10-fold cross-validation, patients are randomly partitioned into ten equal sized parts keeping the class ratio const1ant in each part; nine parts are used for training the classifiers and remaining part is used for testing. This procedure is repeated ten times, resulting each part is tested against the models built using other nine parts. The average of performance measurements of all ten iterations is considered as an unbiased estimate of the whole classification model. We report the performance of the classifiers using standard classification evaluation metrics, including: accuracy, sensitivity (true positive rate, TPR, also called recall), specificity (true negative rate, TNR), false positive rate, false negative rate, precision (Positive Predictive Value, PPV) and F measure (also called F1 score). In Table 2, we show confusion matrix, also called contingency table, which is used to calculate performance measures, and Table 3 values making true positives (TP), false positives (FP), true negative (TN), and false negatives (FN), are shown And the equations to calculate performance measures are presented in Table 3. In addition, we generate

ROC curves, which graphically present the performance of classifiers for each class and calculate the area under the curve (AUC) as a numeric evaluation of ROC curves. Also, we would like to note that even though most of these measures initially defined for binary classification (having only two classes); they are applicable to multiclass classification by following one-vs.-rest approach.

		Actual label		
		Cluster 1	Cluster 2	Cluster 3
Classified as	Cluster 1	a	b	C
	Cluster 2	d	e	F
	Cluster 3	g	h	I

Table 2: Confusion matrix showing the number of patients predicted to be in a class and actual number of patients in that class. As an example value "a" shows the number of patients correctly predicted to be in Cluster 1. And value "b" shows the number of pa

	Cluster 1	Cluster 2	Cluster 3
TP:	a	e	i
FP:	b+c	e+f	h+i
TN:	e+f+h+i	d+g+f+i	d+e+g+h
FN:	d+g	e+h	f+i

Table 3: Shows the definition of basic measures, which are used to calculate performance measures.

(Sensitivity) TPR:	TP/(TP+FN)
(Specificity)TNR:	TN/(TN+FP)
FPR:	FP/(FP+TN) or 1-
FNR:	FN/(FN+TP)
(Precision) PPV:	TP/(TP+FP)
F measure:	2*(PPV*TPR)/(PPV+TPR)

Table 4: Shows the equations to be used to calculate performance measures.

Permutation Test

Finally, to validate the strength of the achieved prediction accuracy, we run a permutation test. For this test we generated 10,000 datasets by randomly shuffling patient labels in our dataset, while keeping the number of patients in each class constant. We run 10-fold cross-validation with RF classification algorithm together with feature selection step on these datasets, in the same way used for the real data in the study. We calculated a p-value by the number of times this validation produced a better accuracy on randomly shuffled dataset divided by 10,000 as seen in Equation 4.

$$Permutation\ test\ P-value = \frac{\#\left(Accuracy_{random} > Accuracy_{original}\right)}{10000}$$

Equation 4: Permutation test

Chapter 5

RESULTS AND DISCUSSION

Exome Data Analysis and Variant Calling

We have obtained an average of 17,640 point variations per patient generated by VarScan2 (Koboldt et al., 2012) and applied a set of filters to select only those that are likely to exhibit an impact on the function and/or the structure of the gene or protein. Since the generation of next-generation sequencing (NGS) data and variant calling involves several error prone steps, filtration of the variant data constitutes a major step in variant analysis. Firstly, we focus only on the somatic (non-inherited) and nonsynonymous (cause a change in the translated amino acid) point mutations because of their perceived impact on disease initiation and progression. Secondly, even though exome sequencing targets only the coding regions of DNA, the exome capture kits often amplify off-target non-coding regions such as intergenic, untranslated and intron regions. Hence, we filter out all the variations outside of the coding region. We analyze the remaining variations by their impact on the function or structure of the resulting protein. Finally, we check the population frequency of remaining variations in Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al., 2001), which is a public achieve for genetic variation developed and hosted by National Center for Biotechnology Information (NCBI). In this step, we filter out the variations that are commonly found in population and hence are not necessarily associated with a disease. Generally, variations with less than 0.05 minor allele frequency (MAF) are considered as phenotype-causing variations and hence are called as mutations.

Classification of Breast Cancers Based On Somatic Mutations

As a prior step before clustering, we applied feature selection by ranking the features (genes) in decreasing order of their variance value and selected top *n* features for clustering. We optimized the size of n to be 854 genes in our clustering method and determined the number of clusters k as explained in "Feature selection and optimization of clustering" section to be 3. Later using the NMF clustering algorithm on our dataset, we stably clustered the samples into 3 groups using the top 854 genes, which have the highest variance values of mutation scores across all the samples. The three groups Cluster 1, 2, and 3 involve 169, 121 and 68 patients, respectively. Refer to Materials and Methods section for more details about the NMF method and to Optimization results section for detailed information on results of the optimization steps.

Unsupervised clustering is the task of grouping a set of samples that have no label information, which results in grouping samples in such a way that samples in the same group are more similar in a specified measure to each other than to those in the other groups. There are several methods trying to achieve this goal such as k-means clustering, hierarchical clustering and expectation maximization (EM) algorithms. However, these methods perform poorly or can not come to a solution when applied to sparse data, as is the case in our study. Therefore, we selected to use NMF because of its proven superior performance when tested on applications that use biological data. (Kim, Seo, Joung, & Kim, 2011; Lawrence et al., 2013; Zheng, Zhang, Ng, Shiu, & Huang, 2011). NMF was introduced in its modern formulation by Lee and Seung (Lee & Seung, 2001) as a method to decompose images.

As a factorization method, NMF algorithm takes our mutation score matrix as the input and decomposes it to two smaller matrices (basis matrix W and coefficient matrix H). The output coefficient matrix (matrix H) is used to make sample cluster assignments. Refer to methods for more details.

In Figure 4 we show a representation of the input data in the mutation score matrix with statistically significant genes sorted by their variance in decreasing order; since the data is extremely sparse the heat map consists of mostly blue cells. In order to make the heat map more readable to human eye, we show the input data in Figure 5 focusing only the top 50 variant genes. As it can be seen, data still represents a very sparse form (most of the cells are colored blue meaning a score of zero) that makes most clustering approaches inapplicable. Figures 6 and 7 are the output matrices from decomposition of the mutation score matrix, which we input to NMF algorithm. Note that multiplication of the two output matrices will approximately yield the input data. In Figure 7, we see the basis matrix (W), which is not used in the scope of this study; however it could serve for clustering purpose of the genes. Figure 6 displays the coefficient matrix (H), where the rows represent the metagenes that are a compact representation of all the genes, and columns represent the patients. We use this matrix to make sample to cluster associations by assigning the samples to the clusters where we observe the highest metagene value, i.e., the dark red color (see methods section for details).

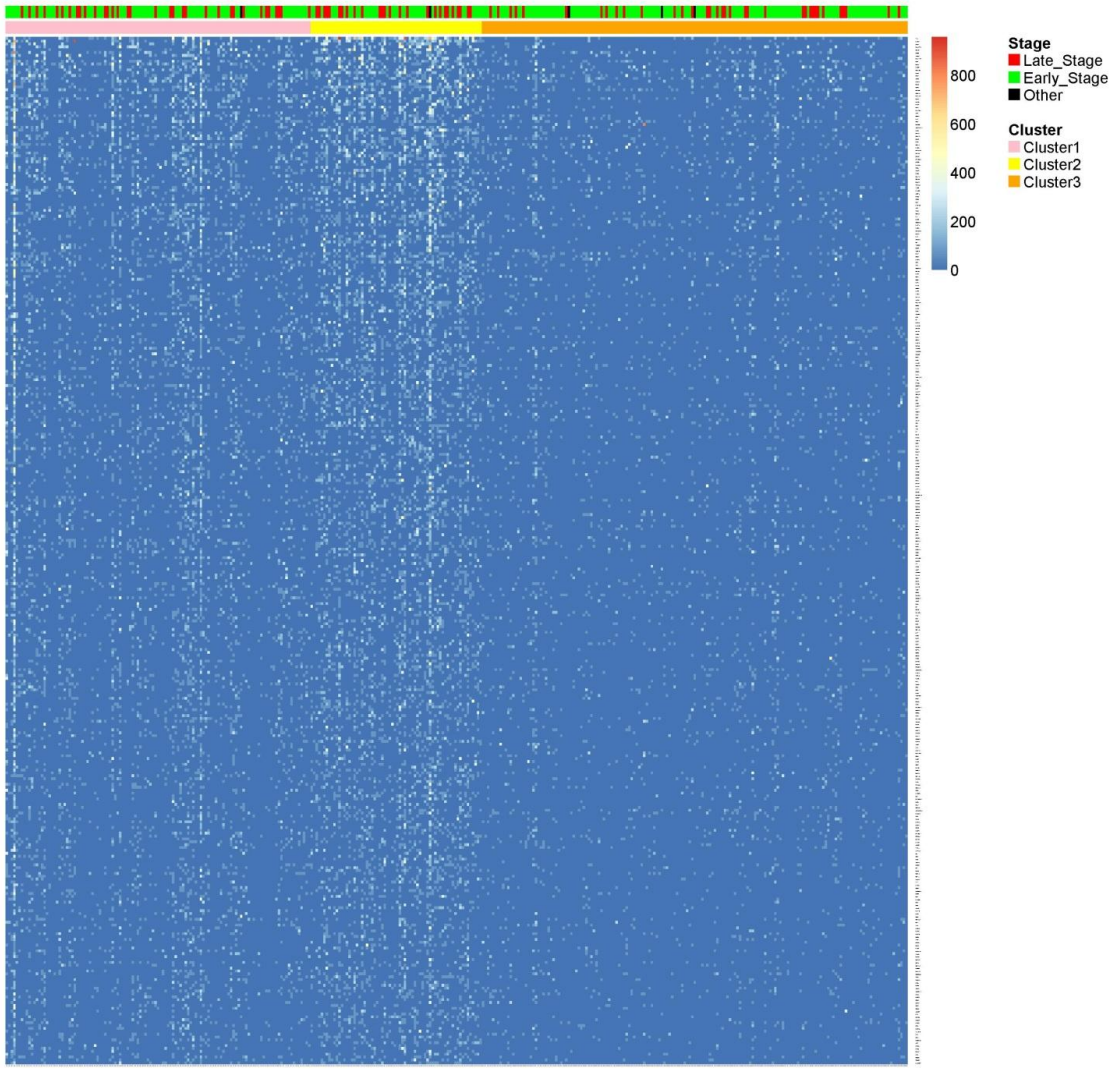


Figure 4: C-scores of 358 significantly mutated genes in late-stage-enriched cluster

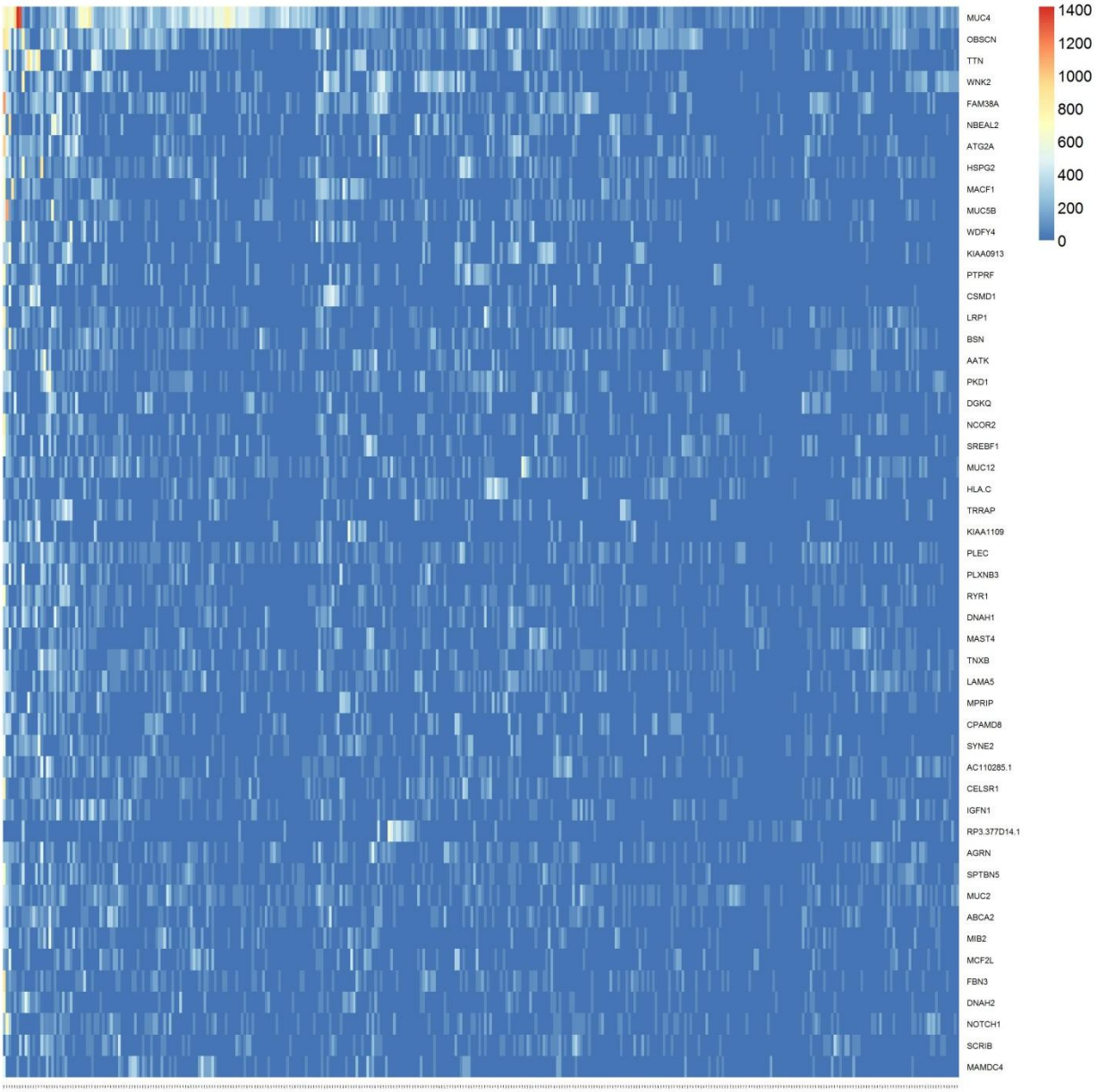


Figure 5: Input matrix with C-scores of the top 50 variant genes. Columns represent patients (358) and rows represent genes.

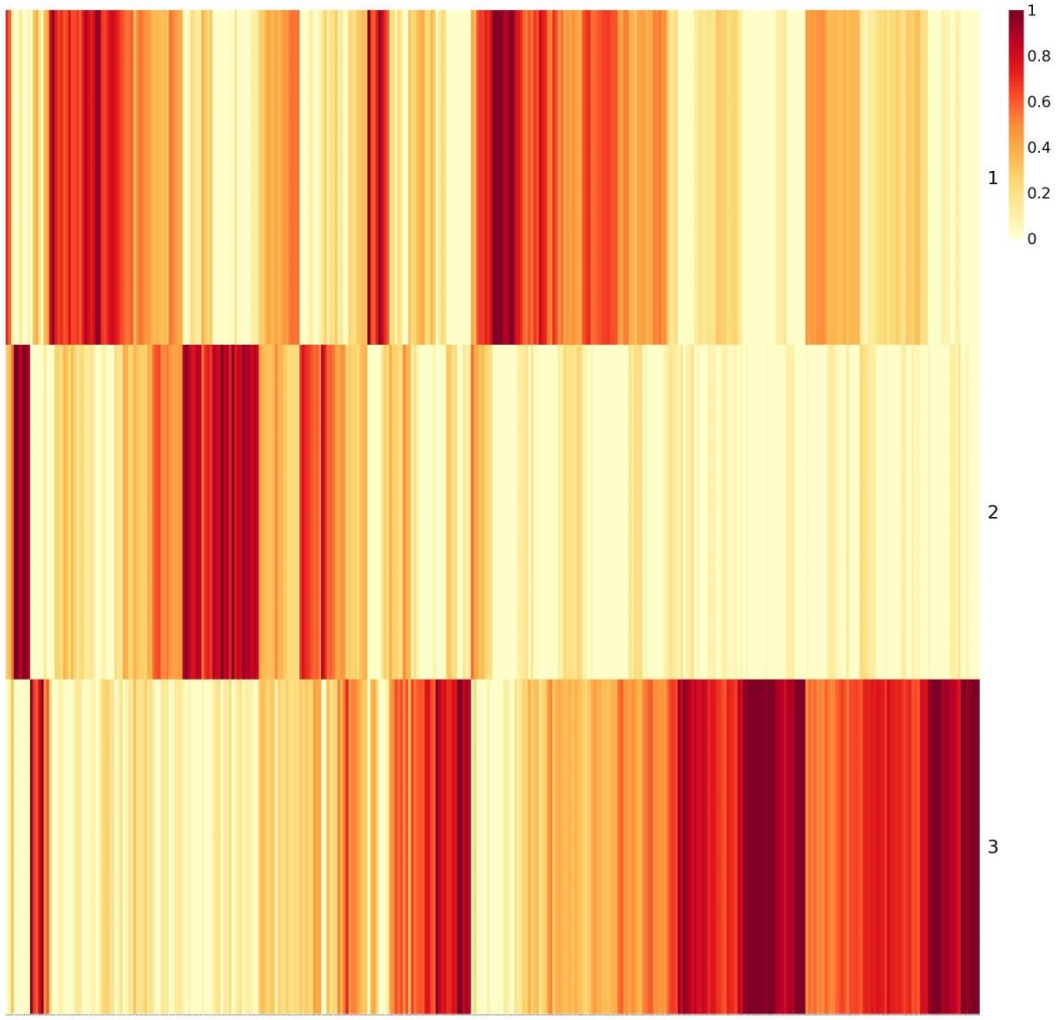


Figure 6: Coefficient matrix (H), 3x358 in size, used for assigning samples to clusters. Columns represent patients and rows represent metagenes.

We generated 3 metagenes that are used to cluster patients into 3 groups. We determined the number of metagenes (rank of clustering) by running the algorithm iteratively over a range of biologically reasonable parameters as explained in the methods section.

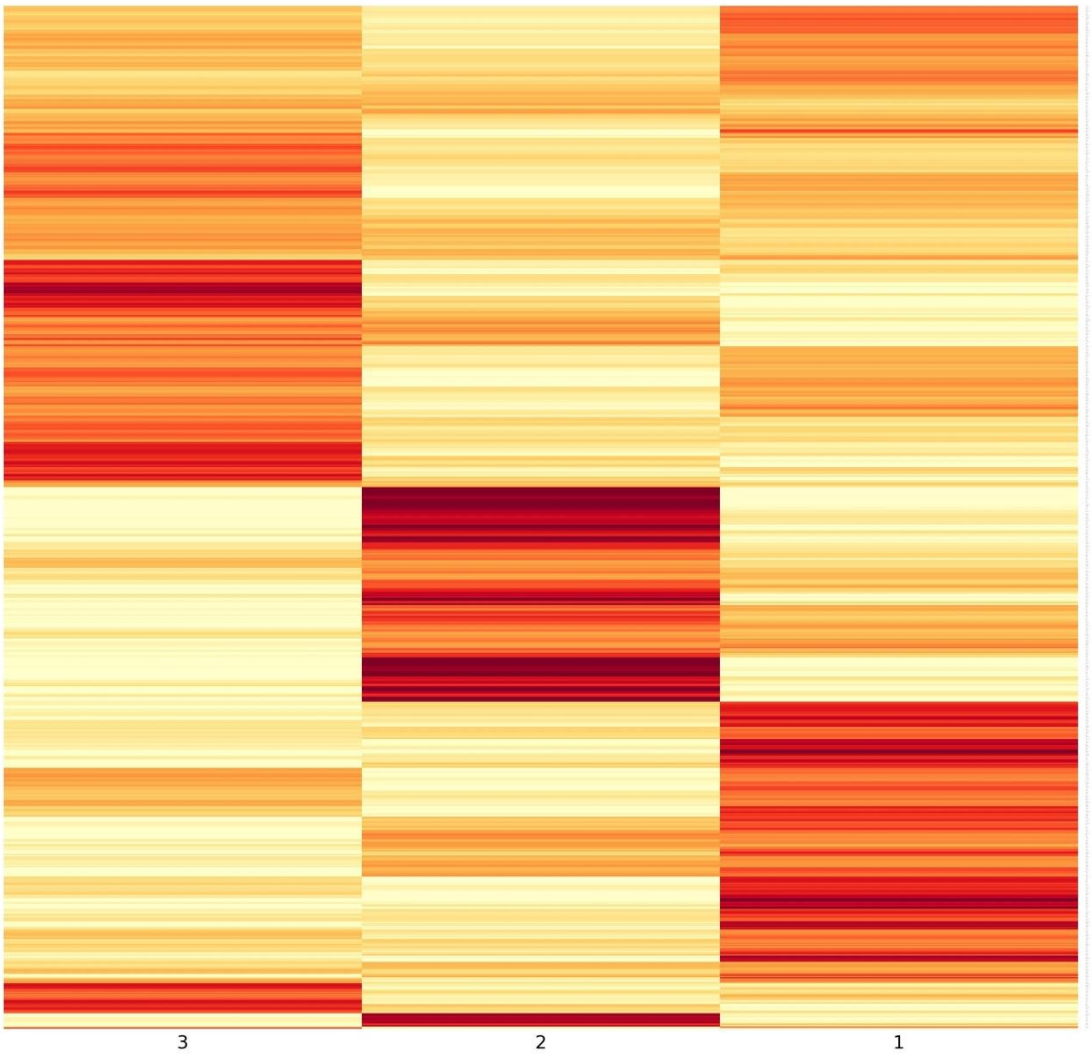


Figure 7: Basis matrix (W), 854x3 in size, clustering the genes.

Figure 8 illustrates the stability of the clustering by displaying the consensus matrix, which was generated after 100 NMF runs using Brunet's (Brunet et al., 2004) approach (explained in methods section). We used the silhouette score of consensus matrix to determine the optimum number of genes and clusters. In an ideal clustering case, we expect to observe values either close to 1 or 0, indicating the probability of two samples being in the same cluster or not, respectively, which displays solid colored blocks. A value of 1 represents the highest probability that two samples are in the same cluster (red blocks) and the value of 0 denotes the opposite (blue blocks). In Fig. 4 it can be seen that the dataset is clearly clustered into three distinct groups.

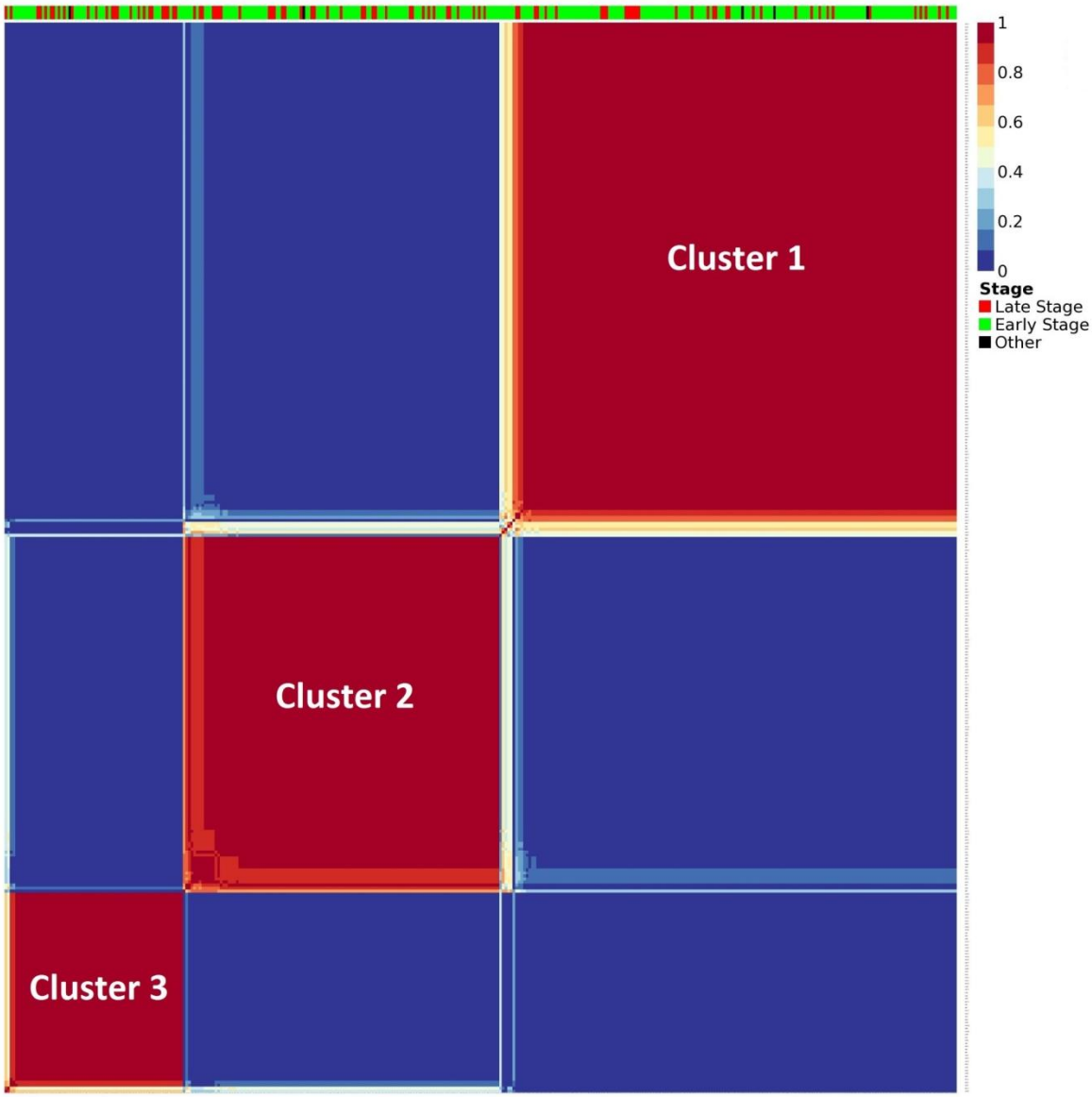
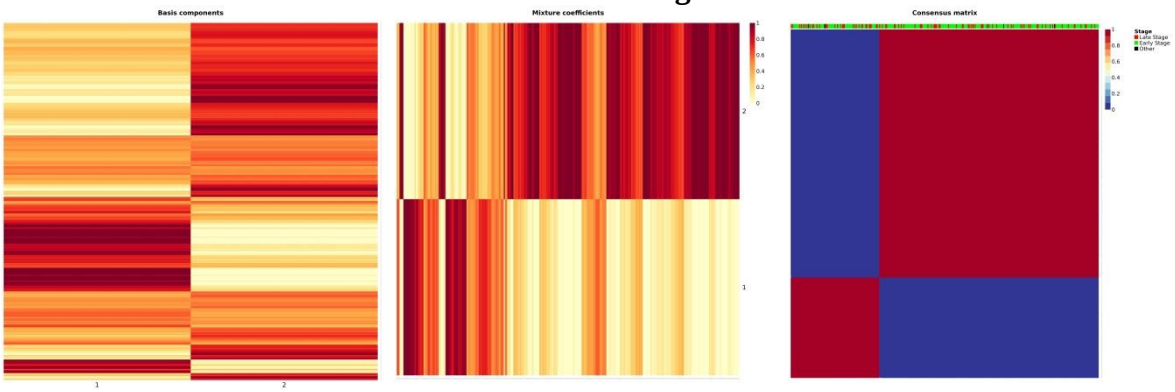


Figure 8: Consensus matrix, 358x358 in size and illustrating the stability of the clustering. In ideal case, all the entries are expected to be either 0 or 1, making solid colored blocks. The bar on top indicates the clinical stage of each patient. Silhouette (consensus) = 0.958

Optimization Results

NMF clustering algorithm takes number of clusters and features as input hence we need to find the optimum number of clusters and genes. Here, we present the results of optimization steps. As explained in the "Feature selection and optimization of clustering" section under "Methods and Materials". We run the clustering algorithm over a range of biologically meaningful parameters that include from 1 to 5 clusters and 10 to 1000 genes. Visual inspection shows the quality of clustering in figures shown in Figure 9 to Figure 12. As a natural consequence of increasing the number of clusters k, the clustering quality decreases. A visual inspection in Basis, Coefficient and Consensus matrices confirms this observation. We observe a decreasing contrast between clusters in Basis and Coefficient matrices and increased block fractures in Consensus matrix, thus decreasing silhouette score, which indicates the decreasing clustering quality. Based on these results, we selected the number of clusters k as 3, and the number of genes n as 854, as optimal. Even though we observe the highest silhouette score in the case k=2, we did not select this parameter because of the infeasible clinical differentiation between clusters.

2 Clusters with 910 genes



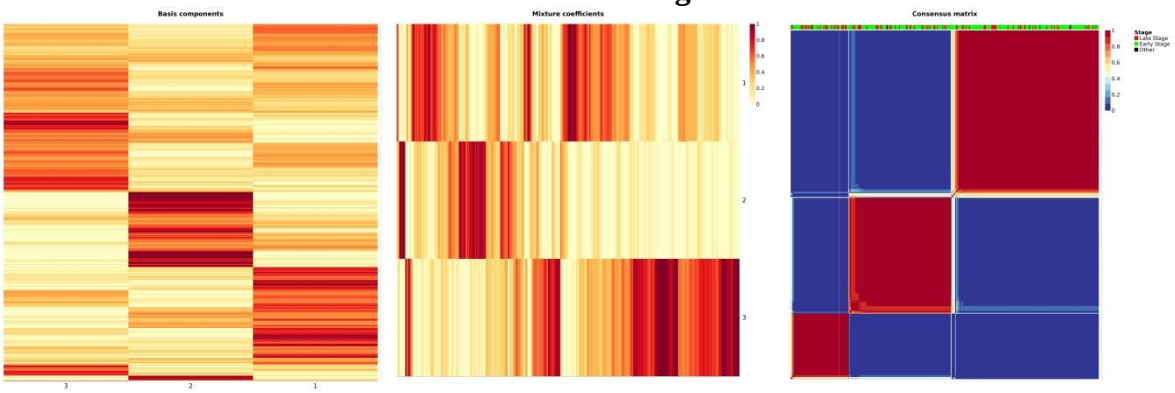
Basis matrix (W)
Dimensions: 910x2

Coefficient matrix (H)
Dimensions: 2x358

Consensus matrix
Dimensions: 358x358
Silhouette score: 99%

Figure 9: Optimal clustering achieved for number of clusters (*k*) selected as 2 using 910 top variant genes. Even though this case achieved the highest silhouette score, the low number of clusters makes the results biologically unexplainable.

3 Clusters with 854 genes



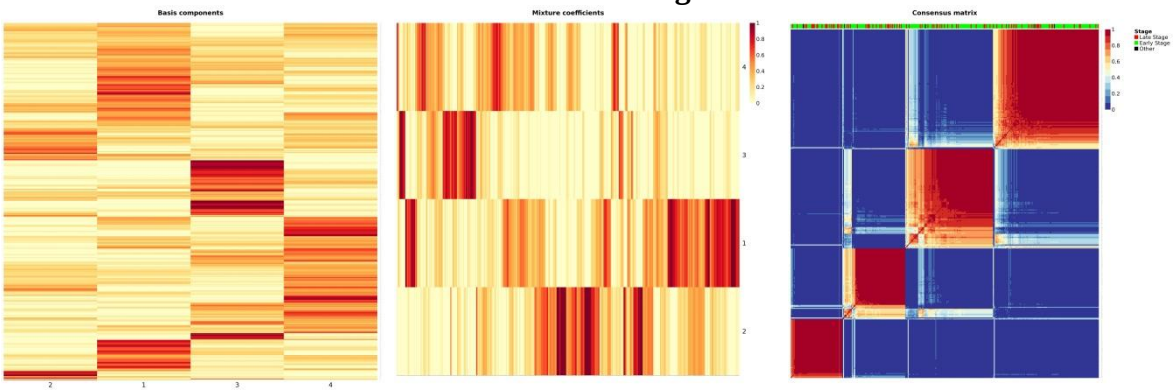
Basis matrix (W)
Dimensions: 854x3

Coefficient matrix (H)
Dimensions: 3x358

Consensus matrix
Dimensions: 358x358
Silhouette score: 96%

Figure 10: Optimal clustering achieved for number of clusters (*k*) selected as 3 using 854 top variant genes. This case is determined to use for further analysis in the project.

4 Clusters with 700 genes



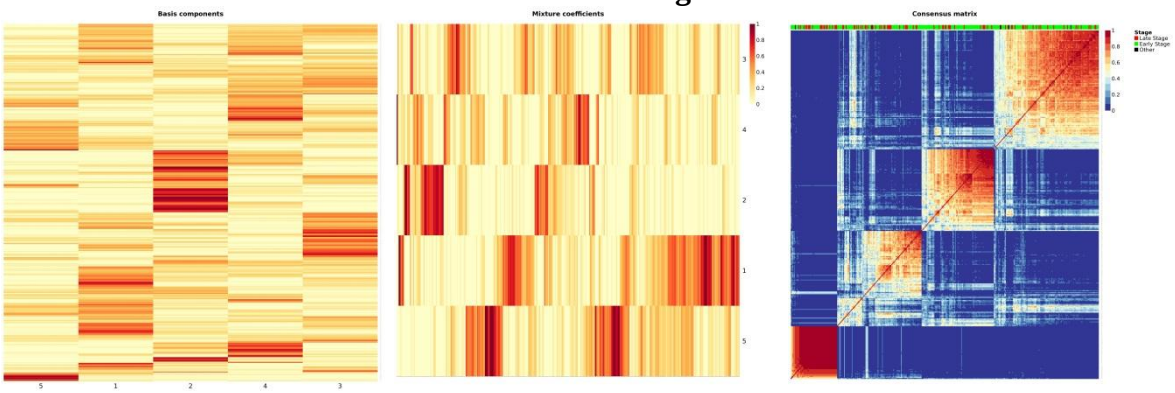
Basis matrix (W)
Dimensions: 700x4

Coefficient matrix (H)
Dimensions: 4x358

Consensus matrix
Dimensions: 358x358
Silhouette score: 84%

Figure 11: Optimal clustering achieved for number of clusters (k) selected as 4 using 700 top variant genes. The deteriorating clustering quality is visible in consensus matrix's heatmap plot and its silhouette score.

5 Clusters with 930 genes



Basis matrix (W)
Dimensions: 930x5

Coefficient matrix (H)
Dimensions: 5x358

Consensus matrix
Dimensions: 358x358
Silhouette score: 52%

Figure 12: Optimal clustering achieved for number of clusters (k) selected as 5 using 930 top variant genes. The deteriorating clustering quality is visible in consensus matrix's heatmap plot and its silhouette score.

Characterization of Discovered Clusters

We investigate the clinical significance of discovered clusters by comparing the BC stage of the patients in each cluster. For this purpose, we analyze the distribution of patients according to their clinical features provided in the TCGA data.

We compare the clusters with a number of clinical features including: ER status, PR status, HER2 status, age at diagnosis, TNBC or non-TNBC, BC stage, and aggregated BC stage. The distribution of patients to groups for two and three cluster cases with p-values generated by Fisher's Exact test can be seen in tables Table 5, Table 6, and Table 7. Even though we observe several significant p-values in the distributions, the mutation load comparisons did not confirm this classification, for this reason we focus our attention only on the BC stage distribution, which was further confirmed by comparison of mutation loads.

	ER Status		PR status		HER2 status	
Cluster	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**
cluster1	18	84	24	78	77	25
cluster2	61	195	91	165	198	58
	P-value	0.2585	P-value	0.0329	P-value	0.7815
Cluster	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**
cluster1	44	125	59	110	137	32
cluster2	27	94	40	81	85	36
cluster3	8	60	16	52	53	15
	P-value:	0.0496	P-value:	0.2331	P-value:	0.1022

Table 5: The distribution of patients to clusters according to their ER, PR and HER2 status with Fisher's Exact test p-values

	Age at diagnosis		TNBC or Non-TNBC		
Cluster	**Mean**		**TNBC**	**Non TNBC**	**Ratio**
cluster1	60.41		27	279	0.0968
cluster2	58.17		132	636	0.2075
	P-value:	0.219	P-value:	0.000399	
Cluster	**Mean**		**TNBC**	**Non TNBC**	**Ratio**
cluster1	57		33	136	0.2426
cluster2	59.66		14	107	0.1308
cluster3	61.79		6	62	0.0968
	P-value:	0.048	P-value:	0.000137442	

Table 6: The distribution of patients to clusters according to their age and TNBC status with Fisher's Exact test p-values. Even though TNBC distribution resulted a significant p-value, it is not used in the project due to comparison of mutation levels of patients contradicts to biological expectation.

	BC Stage				BC Stage		
Cluster	**Stage I**	**Stage II**	**Stage III**	**Stage IV**	**Early Stage***	**Late Stage***	**Ratio***
cluster1	48	153	99	0	201	99	2.0303
cluster2	174	399	177	9	573	186	3.0806
	P-value:	0.0008			P-value:	0.005646	
Cluster	**Stage I**	**Stage II**	**Stage III**	**Stage IV**	**Early Stage***	**Late Stage***	**Ratio***
cluster1	39	92	34	1	131	35	3.7429
cluster2	24	62	32	2	86	34	2.5294
cluster3	11	30	26	0	41	26	1.5769
	P-value:		4.61E-05		P-value:	0.02048	

Table 7: The distribution of patients to clusters according to their BC stage with Fisher's Exact test p-values. This distribution is selected to be further analyzed in the project.

Ratio:	# stage I-II / #stage III-IV
Early stage:	Stage I-II
Late Stage:	Stage III-IV

Table 8: Definition early and late stage breast cancer in the project

In comparison of BC stages we defined two aggregated BC stages as Early and Late stage BC groups, as shown in Table 8 We found that Cluster 1 was dominated by early stage patients while Cluster 3 had much higher proportion of late stage patients compared to Cluster 1 (Fisher's exact test p-value=0.02048, Table 9). As can be seen in Table 9, the number distribution of patients in each cluster with stage ratio (number of early stage patients over late stage patients) for Cluster1 is more than two-fold higher than that of Cluster 3; hence here we call Cluster 1 as the early-stage-enriched cluster, Cluster 2 as the mixed cluster and Cluster 3 as the late-stage-enriched cluster. This

separation of patients by their disease stage indicates that our clustering method can successfully discriminate breast cancer patients by their disease stage using only the somatic mutational profiles of patients from their exome sequencing data.

Cluster	Number of patients [a]	Number of early stage patients [b]	Number of late stage patients [c]	Ratio [d]
Cluster 1	166	131	35	3.74
Cluster 2	120	86	34	2.53
Cluster 3	67	41	26	1.58

Table 9: Distribution of patients in the clusters discovered. P-value= 0.02048
a- Five patients were not included due to their unknown stage information;
b- Sum of stage I and II patients in each cluster;
c- Sum of stage III and IV patients in each cluster;
d- Ratio of the number of early stage patients to the number of late stage patients

Next, we compared the somatic mutation profiles of patients between the early and late-stage-enriched clusters (Cluster 1 vs. Cluster 3). We found that there were 358 genes, which have significantly higher mean mutation scores in the late-stage-enriched cluster (Cluster 3) than in the early-stage-enriched cluster (Cluster 1) (Wilcox rank-sum test, FDR<0.1), but none of the genes have significantly higher mean mutation scores in Cluster 1 than in Cluster 3. This interesting finding indicates that these genes may have accumulated deleterious mutations leading to the progression of breast cancer into advanced disease states. We identified that tumor suppressor genes, APC, BRCA2; and oncogene, MLL are among the 358 genes used in this comparison. Table 10 shows the significant genes that are found to show significantly higher mutation rates in late-stage-enriched cluster.

Gene name	p value	FDR value
TTN	0	0
MACF1	0	0
FSIP2	0	0
DNAH9	0	0
DST	0	0
KIAA1731	0	0
DSP	0	0
VPS13D	4.44E-16	4.74E-14
UBR4	1.55E-15	1.47E-13
C10ORF18	1.89E-15	1.61E-13
SYNE1	2.55E-15	1.98E-13
HERC1	5.44E-15	3.87E-13
CSMD1	2.35E-14	1.55E-12
CHD9	2.93E-14	1.79E-12
KIAA1109	3.26E-14	1.86E-12
XIRP2	4.04E-14	2.16E-12
APC	8.06E-14	4.05E-12
GPR98	1.23E-13	5.86E-12
DOCK9	4.62E-13	2.07E-11
VCAN	5.78E-13	2.47E-11
SYNE2	7.90E-13	3.21E-11
RIF1	1.06E-12	4.10E-11
NOTCH2	1.40E-12	5.21E-11
WDFY4	1.70E-12	6.05E-11
MLL	2.28E-12	7.79E-11
PKHD1	2.89E-12	9.49E-11
AKAP9	5.72E-12	1.81E-10
PHF3	6.79E-12	2.07E-10
STARD9	7.47E-12	2.20E-10
RYR3	7.66E-12	2.18E-10
CEP350	1.15E-11	3.17E-10
LRBA	1.39E-11	3.71E-10
FAT3	1.43E-11	3.69E-10
GOLGA4	2.00E-11	5.02E-10
DNAH2	2.05E-11	5.01E-10
ZNF292	2.24E-11	5.30E-10
WDFY3	2.49E-11	5.74E-10
EYS	4.58E-11	1.03E-09

Gene name	p value	FDR value
NBAS	3.25E-10	5.91E-09
DNAH6	4.83E-10	8.59E-09
CAST	5.00E-10	8.71E-09
CCDC18	5.09E-10	8.70E-09
MBD5	5.39E-10	9.03E-09
KALRN	6.02E-10	9.88E-09
FLNB	6.15E-10	9.91E-09
NAV3	6.24E-10	9.87E-09
BPTF	6.24E-10	9.69E-09
MMS19	6.62E-10	1.01E-08
OTOGL	6.62E-10	9.92E-09
ZFHX4	6.88E-10	1.01E-08
CMYA5	7.51E-10	1.09E-08
MIA3	8.20E-10	1.17E-08
AKAP6	8.62E-10	1.21E-08
SHROOM3	1.23E-09	1.70E-08
DOCK7	2.25E-09	3.06E-08
DYNC2H1	2.37E-09	3.17E-08
HECTD1	2.40E-09	3.15E-08
PTPRQ	2.83E-09	3.66E-08
WHSC1	2.83E-09	3.61E-08
TET1	2.97E-09	3.73E-08
NBEA	3.07E-09	3.80E-08
COL4A3	4.80E-09	5.85E-08
LPHN2	7.23E-09	8.69E-08
BRCA2	8.91E-09	1.06E-07
MLPH	1.06E-08	1.24E-07
HIVEP1	1.08E-08	1.25E-07
C12ORF35	1.08E-08	1.23E-07
QSER1	1.13E-08	1.27E-07
BIRC6	1.19E-08	1.32E-07
FAT1	1.31E-08	1.44E-07
ZDBF2	1.72E-08	1.85E-07
NUMA1	1.78E-08	1.90E-07
N4BP2	1.80E-08	1.90E-07
GCN1L1	2.05E-08	2.14E-07
KCNMA1	4.42E-08	4.55E-07
HUWE1	4.42E-08	4.49E-07

SACS	8.23E-11	1.80E-09	TTC3	4.88E-08	4.90E-07
FAT4	1.07E-10	2.28E-09	MYCBP2	5.03E-08	4.99E-07
PRRC2C	1.58E-10	3.29E-09	CEP250	7.33E-08	7.20E-07
ANK2	1.65E-10	3.36E-09	HYDIN	1.36E-07	1.32E-06
LRP2	1.68E-10	3.33E-09	RP1	1.40E-07	1.34E-06
LIMCH1	1.71E-10	3.32E-09	MDN1	1.42E-07	1.34E-06
FRY	1.75E-10	3.33E-09	EPG5	1.78E-07	1.67E-06
CENPF	1.95E-10	3.61E-09	PLB1	1.87E-07	1.74E-06
PMEL	1.87E-07	1.72E-06	LPHN3	1.07E-05	6.54E-05
KLKB1	1.87E-07	1.70E-06	MLL3	1.14E-05	6.93E-05
MGA	1.95E-07	1.75E-06	TG	1.18E-05	7.13E-05
RAPGEF2	2.13E-07	1.90E-06	XKR6	1.23E-05	7.39E-05
ANKRD12	2.18E-07	1.92E-06	NUP98	1.23E-05	7.34E-05
LMO7	2.24E-07	1.95E-06	FRAS1	1.25E-05	7.44E-05
LAMA3	2.45E-07	2.11E-06	ASXL3	1.35E-05	7.94E-05
PRKDC	2.84E-07	2.42E-06	CSMD2	1.48E-05	8.67E-05
BRPF1	3.02E-07	2.56E-06	C12ORF51	1.57E-05	9.14E-05
ADARB1	4.63E-07	3.87E-06	HTT	1.57E-05	9.08E-05
FHAD1	4.66E-07	3.86E-06	SZT2	2.42E-05	1.39E-04
WNK1	5.29E-07	4.34E-06	TLN2	2.54E-05	1.45E-04
TNRC6B	5.32E-07	4.33E-06	DCHS1	2.67E-05	1.51E-04
HEATR5A	5.57E-07	4.49E-06	RERE	2.72E-05	1.53E-04
ODZ2	5.97E-07	4.77E-06	NCAPG2	2.78E-05	1.55E-04
MYO18B	6.52E-07	5.16E-06	TTC28	3.05E-05	1.69E-04
USP34	6.77E-07	5.31E-06	MAGI1	3.20E-05	1.76E-04
PDZD2	6.96E-07	5.41E-06	DMXL1	3.22E-05	1.76E-04
CASC5	7.38E-07	5.68E-06	ARID1A	3.24E-05	1.76E-04
ALMS1	7.88E-07	6.01E-06	DNAH1	3.41E-05	1.85E-04
SPEN	8.18E-07	6.18E-06	MAPKBP1	3.49E-05	1.88E-04
EP300	8.38E-07	6.28E-06	DNAH10	3.63E-05	1.94E-04
LOXHD1	1.21E-06	9.01E-06	NAV2	3.87E-05	2.06E-04
C11ORF41	1.38E-06	1.02E-05	ANKRD11	4.00E-05	2.11E-04
ARNTL2	1.41E-06	1.03E-05	ATP13A1	4.03E-05	2.11E-04
TRRAP	1.48E-06	1.07E-05	CIT	4.10E-05	2.14E-04
SEC31A	1.49E-06	1.07E-05	ABCA13	4.53E-05	2.35E-04
CCDC136	1.87E-06	1.33E-05	ANK1	4.80E-05	2.47E-04
KIAA0226	2.26E-06	1.59E-05	TACC2	5.42E-05	2.77E-04
EHBP1	2.80E-06	1.96E-05	FAT2	6.06E-05	3.08E-04
JMJD1C	2.80E-06	1.94E-05	SASH1	6.68E-05	3.38E-04

EDC4	2.87E-06	1.97E-05	SPTBN1	6.73E-05	3.38E-04
AKAP12	3.18E-06	2.17E-05	TCOF1	7.68E-05	3.83E-04
COL4A5	3.62E-06	2.45E-05	LAMB2	1.10E-04	5.49E-04
SYNPO2	3.84E-06	2.58E-05	TEP1	1.17E-04	5.76E-04
CHD3	4.07E-06	2.71E-05	GRIK5	1.17E-04	5.74E-04
SPTAN1	4.34E-06	2.87E-05	TRPM3	1.20E-04	5.88E-04
SRRM2	5.62E-06	3.69E-05	MED12	1.24E-04	6.00E-04
COL6A3	6.09E-06	3.97E-05	DIP2C	1.29E-04	6.20E-04
CEP164	6.96E-06	4.50E-05	ZNF407	1.30E-04	6.22E-04
DNAH14	8.36E-06	5.37E-05	SALL2	1.31E-04	6.26E-04
PUM1	8.53E-06	5.43E-05	RYR1	1.76E-04	8.37E-04
DEPDC5	8.70E-06	5.50E-05	CACNA1A	2.22E-04	1.05E-03
SETD5	9.49E-06	5.96E-05	TRAPPC9	2.41E-04	1.13E-03
DOCK5	9.70E-06	6.05E-05	DYNC1H1	2.68E-04	1.25E-03
IPO5	9.70E-06	6.00E-05	SF1	2.83E-04	1.31E-03
CARD10	2.99E-04	1.38E-03	CDK20	1.76E-03	6.49E-03
CEP128	3.65E-04	1.68E-03	ARHGEF40	1.77E-03	6.52E-03
ROBO3	3.81E-04	1.74E-03	RPS6KA4	1.84E-03	6.76E-03
PCNX	4.25E-04	1.93E-03	UROC1	1.99E-03	7.25E-03
PLEKHH1	4.26E-04	1.93E-03	NLRC5	2.24E-03	8.14E-03
ERGIC3	4.30E-04	1.93E-03	SSC5D	2.27E-03	8.20E-03
COL17A1	4.31E-04	1.93E-03	PKD1L2	2.27E-03	8.19E-03
HERC2	4.50E-04	2.00E-03	MCF2L	2.36E-03	8.46E-03
CDC27	4.54E-04	2.01E-03	ODZ4	2.89E-03	1.03E-02
COL7A1	4.63E-04	2.04E-03	MST1R	2.96E-03	1.05E-02
CD44	4.75E-04	2.08E-03	KIAA1967	3.05E-03	1.08E-02
FANCM	5.60E-04	2.44E-03	GBGT1	3.05E-03	1.08E-02
MTUS2	5.71E-04	2.48E-03	CUX2	3.08E-03	1.08E-02
SPATA13	5.73E-04	2.47E-03	LRRC16B	3.26E-03	1.14E-02
ACACB	5.87E-04	2.52E-03	DNAH17	3.45E-03	1.20E-02
DNHD1	6.66E-04	2.84E-03	MYH9	3.58E-03	1.24E-02
EPB41L1	6.84E-04	2.91E-03	SON	3.67E-03	1.27E-02
TLN1	7.55E-04	3.19E-03	PLOD3	3.74E-03	1.29E-02
RBM33	7.61E-04	3.20E-03	STK36	3.97E-03	1.36E-02
FANCA	7.79E-04	3.26E-03	LTBP3	4.01E-03	1.37E-02
TET3	7.86E-04	3.27E-03	COL4A4	4.25E-03	1.45E-02
RNF213	8.07E-04	3.35E-03	RLTPR	4.74E-03	1.61E-02
CAMTA1	8.23E-04	3.40E-03	MICAL1	4.80E-03	1.62E-02
DLEC1	8.26E-04	3.39E-03	RUSC2	4.80E-03	1.62E-02

MAP1A	8.28E-04	3.38E-03	PXDN	4.86E-03	1.63E-02
CCDC108	8.38E-04	3.41E-03	STARD8	5.02E-03	1.67E-02
CABIN1	8.39E-04	3.40E-03	AD000671.1	5.09E-03	1.69E-02
PCDH7	8.94E-04	3.60E-03	SEMA5B	5.27E-03	1.74E-02
NACA	9.62E-04	3.86E-03	ATP11A	5.30E-03	1.75E-02
KIAA0913	1.00E-03	4.01E-03	P2RX2	5.42E-03	1.78E-02
LRRK1	1.02E-03	4.04E-03	IGFN1	5.44E-03	1.78E-02
PIK3CD	1.07E-03	4.25E-03	MYO7A	5.57E-03	1.82E-02
MLLT4	1.07E-03	4.23E-03	TRPM5	5.75E-03	1.87E-02
RALGAPA2	1.12E-03	4.40E-03	ESPL1	5.76E-03	1.86E-02
COL5A1	1.27E-03	4.95E-03	KIAA1274	5.97E-03	1.92E-02
L1CAM	1.28E-03	4.97E-03	ZNF536	6.39E-03	2.05E-02
CCDC165	1.29E-03	4.97E-03	AHNAK2	6.41E-03	2.05E-02
DPP9	1.31E-03	5.05E-03	KIF26B	6.50E-03	2.07E-02
DOCK4	1.32E-03	5.04E-03	ZFHX3	6.51E-03	2.07E-02
COL6A6	1.42E-03	5.42E-03	MXRA5	6.61E-03	2.09E-02
TECTA	1.44E-03	5.46E-03	NCOR2	6.69E-03	2.11E-02
FAM65A	1.59E-03	6.02E-03	RNF207	6.88E-03	2.16E-02
SPEG	1.61E-03	6.04E-03	EHMT1	6.98E-03	2.18E-02
AC136932.2	1.62E-03	6.06E-03	ARAP3	7.01E-03	2.19E-02
LRP4	1.70E-03	6.34E-03	MPRIP	7.17E-03	2.23E-02
PCNT	1.71E-03	6.35E-03	ADAP1	7.88E-03	2.44E-02
KDM6B	7.99E-03	2.46E-02	FER1L5	2.20E-02	5.81E-02
CUL9	8.75E-03	2.69E-02	SHANK3	2.23E-02	5.88E-02
COL27A1	8.95E-03	2.74E-02	CELSR3	1.94E-02	5.27E-02
VWF	8.95E-03	2.73E-02	ATP8B3	1.94E-02	5.27E-02
GTF2IRD1	9.44E-03	2.87E-02	C20ORF132	1.94E-02	5.26E-02
ADAMTS10	9.62E-03	2.91E-02	ZNF335	2.02E-02	5.44E-02
CPAMD8	9.83E-03	2.97E-02	CDH23	2.03E-02	5.45E-02
NAV1	1.00E-02	3.01E-02	ZNF574	2.08E-02	5.58E-02
OGG1	1.01E-02	3.04E-02	SCAP	2.12E-02	5.65E-02
TBC1D9B	1.04E-02	3.09E-02	ARID1B	2.12E-02	5.65E-02
SFXN5	1.11E-02	3.29E-02	NUP210	2.18E-02	5.79E-02
MUC16	1.13E-02	3.36E-02	CLU	2.36E-02	6.21E-02
ITGA2B	1.16E-02	3.42E-02	GUCY2D	2.42E-02	6.35E-02
NPHP4	1.19E-02	3.50E-02	BAI2	2.46E-02	6.42E-02
C9ORF79	1.19E-02	3.49E-02	OSBP2	2.53E-02	6.59E-02
GPR123	1.20E-02	3.51E-02	INPPL1	2.58E-02	6.71E-02
SREBF2	1.24E-02	3.60E-02	EML3	2.66E-02	6.89E-02

DCHS2	1.35E-02	3.91E-02	CASZ1	2.86E-02	7.29E-02
MAN2B2	1.43E-02	4.14E-02	PTPRS	2.86E-02	7.28E-02
THADA	1.45E-02	4.18E-02	ABCA2	2.95E-02	7.47E-02
COL22A1	1.47E-02	4.24E-02	DIP2A	3.03E-02	7.65E-02
AGAP2	1.48E-02	4.24E-02	PCDH17	3.04E-02	7.66E-02
SHANK1	1.49E-02	4.26E-02	SLC12A3	3.28E-02	8.23E-02
FADS3	1.51E-02	4.29E-02	SLC9A5	3.36E-02	8.42E-02
DLK2	1.56E-02	4.41E-02	KCNQ4	3.44E-02	8.59E-02
DYSF	1.56E-02	4.41E-02	FLII	3.45E-02	8.59E-02
PXDNL	1.58E-02	4.44E-02	RGS14	3.49E-02	8.67E-02
URB1	1.60E-02	4.50E-02	PROC	3.52E-02	8.71E-02
HEATR2	1.64E-02	4.60E-02	FKBP10	3.55E-02	8.76E-02
IRAK1	1.66E-02	4.63E-02	B4GALNT4	3.57E-02	8.79E-02
EHBP1L1	1.71E-02	4.77E-02	RGS3	3.61E-02	8.87E-02
D2HGDH	1.72E-02	4.76E-02	COL2A1	3.68E-02	9.00E-02
COL16A1	1.73E-02	4.79E-02	ZNF541	3.72E-02	9.07E-02
SCNN1D	1.79E-02	4.93E-02	ALPK3	3.80E-02	9.24E-02
ZC3H18	1.83E-02	5.03E-02	ARHGAP33	3.91E-02	9.48E-02
BRD1	1.87E-02	5.12E-02	TNS3	3.94E-02	9.53E-02
IFFO1	1.91E-02	5.21E-02	EPHA1	3.96E-02	9.55E-02
ITGB5	2.75E-02	7.09E-02	INTS1	4.00E-02	9.61E-02
RHBDF1	2.80E-02	7.20E-02	CCDC88C	4.00E-02	9.60E-02
AEBP1	2.81E-02	7.20E-02	TPO	4.17E-02	9.98E-02
GATAD2A	2.82E-02	7.20E-02	PTCH2	4.20E-02	1.00E-01

Table 10: Significant genes that show higher mutation rates in late-stage- enriched cluster (cluster 3)

We stratified these 358 genes into different gene families using the Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) tool as shown in Table 11. We observe that a significant proportion of the genes belong to transcription factor and protein kinase gene families, which are well known to be related to the progression of BC (Adeyinka, Nui, Cherlet, & Snell, 2002; Subramanian et al., 2005) Table 12 shows the assignment of these genes to functionally distinct gene families.

GSEA gene families	Cytokines /growth factors	Transcription factors	Homeodomain proteins	Cell differentiation markers	Protein kinases	Translocated cancer genes	Oncogenes	Tumor suppressors
Tumor suppressors	0	1	0	0	0	1	0	4
Oncogenes	0	3	0	0	0	11	12	
Translocated cancer genes	0	4	0	0	0	12		
Protein kinases	0	0	0	1	16			
Cell differentiation markers	0	0	0	4				
Homeodomain proteins	0	3	3					
Transcription factors	0	25						
Cytokines and growth factors	3							

Table 11: GSEA classification of 358 genes that have significantly higher mean mutation scores in cluster 3 compared to cluster 1. Note that some of the genes in our gene list are not found in any GSEA gene family.

Transcription factors	Protein kinases	Translocated cancer genes	Oncogenes	Cell differentiation markers	Tumor suppressors	Homeodomain proteins	Cytokines and growth factors
ARID1B	ALPK3	AKAP9	AKAP9	CD44	APC	CUX2	LTBP3
BPTF	CDK20	CASC5	CASC5	ITGA2B	BRCA2	ZFHX3	SEMA5B
BRD1	CIT	EP300	MLL	L1CAM	EP300	ZFHX4	TG
BRPF1	EPHA1	MLL	MLLT4	MST1R	FANCA		
CASZ1	GUCY2D	MLLT4	MYH9				
CHD3	IRAK1	MYH9	NACA				
CUX2	KALRN	NACA	NOTCH2				
EP300	LRRK1	NUMA1	NUMA1				
HIVEP1	MST1R	NUP98	NUP98				
LMO7	PRKDC	RNF213	RNF213				
MED12	RPS6KA4	TET1	TET1				
MGA	SPEG	WHSC1	WHSC1				
MLL	STK36						
MLLT4	TRRAP						
NCOR2	TTN						
PHF3	WNK1						
RERE							
SALL2							
SF1							
SPEN							
SREBF2							
UBR4							
WHSC1							
ZFHX3							
ZFHX4							

Table 12: Distribution of genes to functionally distinct gene families by GSEA

Network Analysis of Differentially Mutated Genes

We carried out the network analysis of the top 25 highly mutated genes (Table 2) in the late-stage-enriched cluster compared to the early-stage-enriched cluster patients, to understand the functional relationship among these genes. The network in Figure 13 generated using the Ingenuity Pathway Analysis (IPA) program shows several interaction hubs, where the genes highlighted in purple color are highly mutated in the late stage cluster patients. Most of the genes in our list interact with the central hub protein, UBC, which is expected because most of the proteins (especially the unneeded or damaged ones) are ubiquitinated before proteosomal degradation. It has been known that ubiquitin-proteasome system regulates the degradation of a number of cancer-associated genes (Adams, 2003). APC (adenomatous polyposis coli) is another key tumor suppressor seen in this network that acts as an antagonist of the Wnt signaling pathway, with a number of roles in cancer development and progression such as cell migration, adhesion, apoptosis, etc. The role of APC mutations in breast cancers has been well documented in the literature (Furuuchi et al., 2000). It is noteworthy to mention two transcriptional regulator genes in our list, NOTCH2 and KMT2A (MLL). NOTCH2 is a key regulator of Akt, and its role is well documented in several cancers including in apoptosis, proliferation and epithelial-mesenchymal transition (EMT) pathway (Güngör et al., 2011). Several somatic mutations in NOTCH2 are also associated with different cancers in COSMIC database (Forbes et al., 2014). MLL is a transcriptional regulator and an oncogene with a variety of roles in cell proliferation and apoptosis (Won Jeong, Chodankar, Purcell, Bittencourt, & Stallcup, 2012).

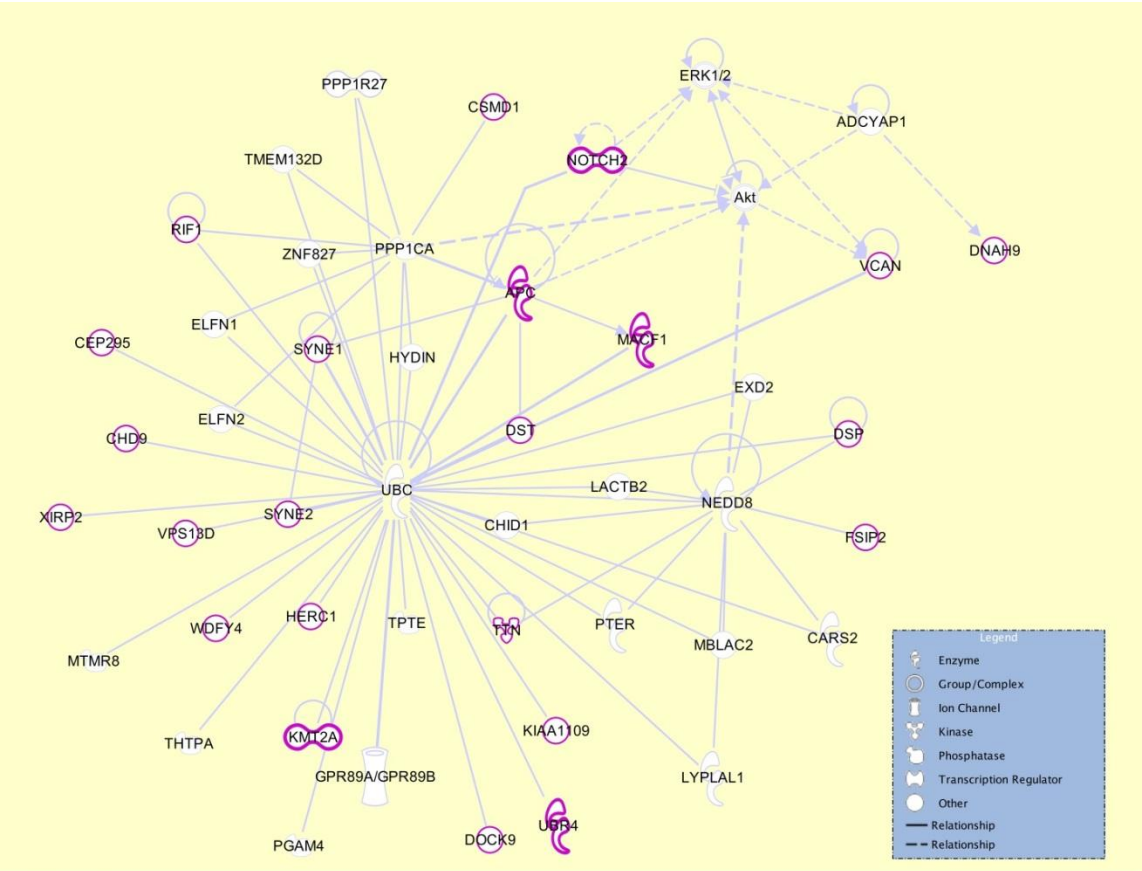


Figure 13: Interaction network analysis of the top 25 genes showing the highest mutation load in the late-stage-enriched cluster compared to the early-stage-enriched cluster of patients.

Class Prediction of Breast Cancers Based On Somatic Mutations

Using the aforementioned BC clusters, we developed a classification model to see how accurate we can predict them based on somatic mutations. With this model, we can identify the stage of a given patient, given the mutation profile of a patient. As an example; if the model predicts a new patient to be in the Cluster3, than we can expect this patient to be in late stage with certain genes be more likely to carry higher mutation loads.

We labeled each patient with its assigned cluster and tested five popular machine learning (ML) algorithms; Random Forest (RF) (Breiman, 2001), Support Vector Machine (SVM) (Platt, 1998) , C4.5 (Salzberg, 1994), Naïve Bayes (Rish, 2001) , and k-Nearest Neighbor(KNN) (Stevens et al., 1967) to find the most appropriate algorithm for our dataset.

We used a 10-fold cross-validation for evaluation of classifier performances. In each loop of the 10-fold cross validation, after withdrawal of the test set, we did feature selection using the information gain feature selection method (Mitchell, 1997) and selected the top 500 genes, which provide the highest information gain based on the training set. Therefore, in total, we selected 10 sets of 500 genes in the 10-fold cross validation. Out of the aforementioned ML algorithms, we selected to further use the RF method in this study as it achieved the best 10-fold cross-validation accuracy with 70.86 %. We believe that the sparseness of the data along with the low sample to feature ratio are the reasons behind this moderate accuracy.

Also we observe that SVM algorithms achieved a very close accuracy but with a loss in TPR, FPR and F measure. And KNN method yielded the worst accuracy of all the methods we used. Table 13 shows the performance measures of each ML algorithm.

Classifier	Accuracy	TPR	FPR	TNR	FNR	F measure
RF	70.86	0.58	0.19	0.81	0.42	0.59
SVM	69.16	0.49	0.16	0.84	0.51	0.53
J48 (C4.5)	60.11	0.47	0.26	0.74	0.53	0.47
Naïve Bayes	57.24	0.45	0.29	0.71	0.55	0.44
k-NN	49.17	0.25	0.16	0.84	0.75	0.31

Table 13: 10-fold cross-validation performance results of five classifiers.

Figure 14 shows the receiver operating characteristic (ROC) curves for each class that illustrate the relationship between TPR (sensitivity) and FPR (1-specificity) for each class. In the perfect case, an ROC curve goes straight up on the Y-axis and then to the right parallel to the X-axis; thus maximizing the area under the curve (AUC). An AUC close to 1 indicates that the classifier is predicting with maximum TP and minimum FP. We calculated the AUC for clusters 1, 2 and 3 (used interchangeably as class in this section) as 0.88, 0.8 and 0.95, respectively, indicating that the classification model can better differentiate the late stage patients against the remaining patients.

We also used a permutation test, by running the same class prediction procedure with RF on 10,000 randomly labeled datasets and none of the 10-fold cross-validations gave us a better accuracy, yielding a very significant p-value (p-value$<10^{-4}$) (see methods for more details). This supports the robustness of our model and the predication accuracy.

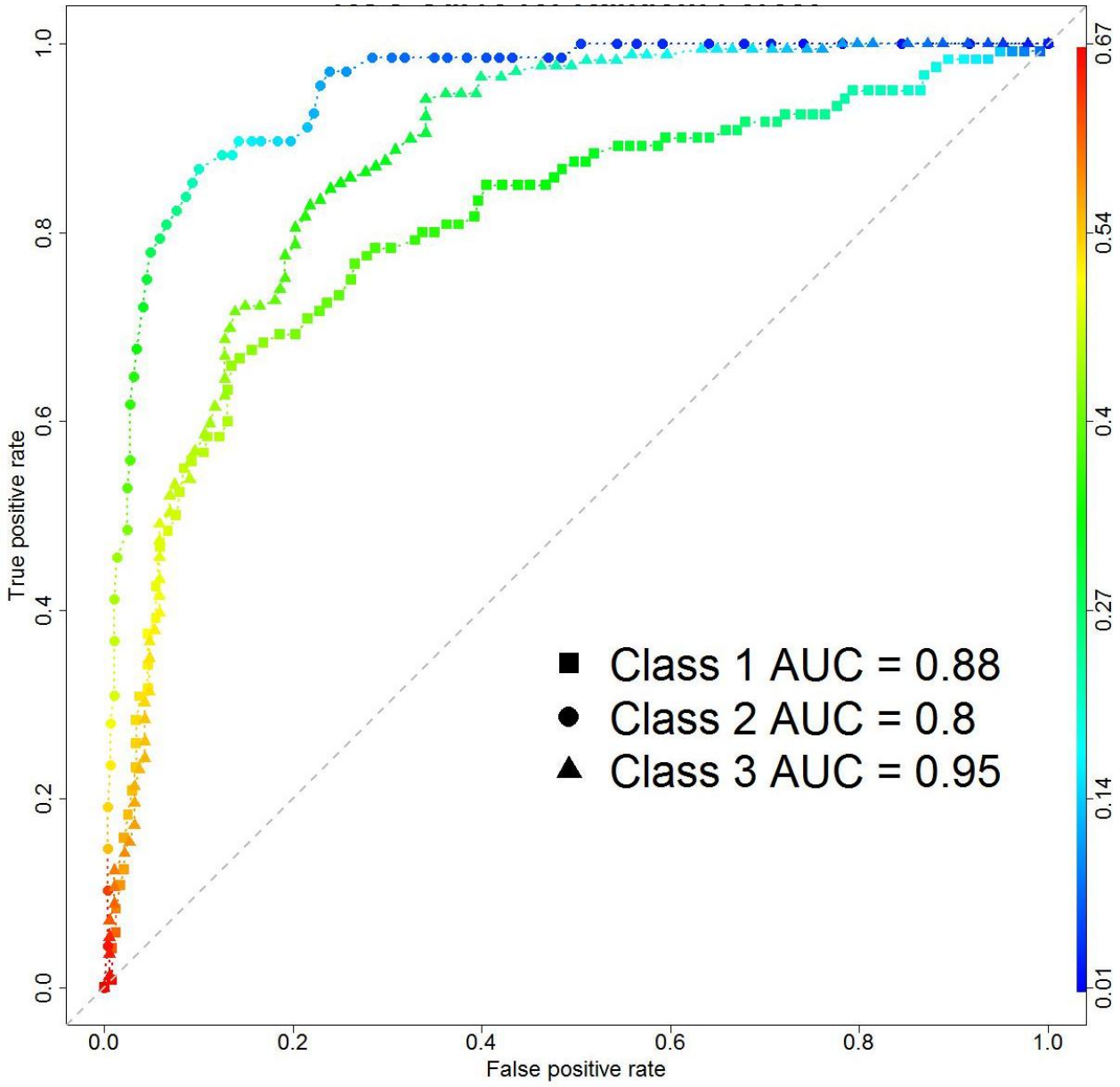


Figure 14: ROC curves showing the relationship between TPR (sensitivity) and FPR (1-specificity) for each class.

Chapter 6

CONCLUSIONS

Breast cancer is a highly heterogeneous disease; therefore, accurate classification of BCs is an important step towards making accurate treatment decisions. Next generation sequencing has opened up new venues to better understand the genomic background of BC at the molecular level. In this study, we developed a novel BC classification system that solely uses somatic mutational profiles of BC patients, generated by whole exome sequencing, to identify clinically distinguishable subgroups together with a class prediction model.

We used the TCGA breast cancer somatic mutation dataset including 358 patients and applied necessary filtration to the reported variations. Following, we used NMF clustering method to discover subgroups in the dataset, which yielded 3 clustered groups of patients. We investigated the clinical significance of discovered clusters by comparing the BC stage of the patients in the clusters and found that there exists a significant separation of patients according to their disease stage; hence we named Cluster 1 as early-stage-enriched and Cluster 3 as late-stage-enriched. Then we compared the mean mutation scores of early and late-stage-enriched clusters and found that late-stage-enriched cluster patients carry a significantly higher rate of mutations in 358 genes. We also identified important networks, biological functions and pathways regulated by these genes. Finally, we used RF classification algorithms to develop a classification model, to make cluster predictions for unknown BC patients, which can provide insights about the disease stage and significantly mutated genes.

In conclusion, this study demonstrates that clinically distinguishable breast cancer subtypes can be identified solely based on somatic mutation profile data from breast cancer patients. Further, our classification model can be used to predict the unknown subtypes of breast cancers, given the somatic mutation profile of a patient. This generic methodology can also be applied to classify and predict other cancer types.

Future Directions

Our work presented here attempts to demonstrate that somatic mutation data alone could be used to classify clinically distinguishable breast cancers. However, additional types of structural data such as copy number variations (CNVs) and insertions/deletions (indels) could also be added to augment the classification accuracy. The main limitation of this work is the lack of clinical data on patients such as survival information or molecular subtype, which directly limit our ability to correlate the identified clusters to clinical features. Hence in our future work, we plan to use a bigger cohort of patients from cancer registries that document longitudinal history of clinical parameters. It will also help us to improve the prediction accuracy of the supervised learning models, which would allow us to build an accurate prediction tool that can be offered to the research community as a web server. As more genomic data becomes available, we expect to use other types of structural data (indels, CNVs, etc.) for improving the prediction models and developing a better breast cancer classification system.

APPENDIX

Here we present a selected set of programs that were developed to carry out the main tasks in this project

Building The Main Data Structure

Here, we present the Python code developed to build the main data structure. Basically code collects each patient's mutations, applies some filters and compiles all the data in a big table, which is later used in NMF clustering.

```
#!/usr/bin/python
import sys,os,argparse,time,glob,csv
import functions
start_time=time.time()

##########################
skip_header_lines=functions.skip_header_lines
get_patients_list=functions.get_patients_list

CADD_raw=5           #range: -inf,+inf
CADD_raw_rankscore=6  #range: [0,1]
CADD_phred=7          #range: [0,1]

#Config-------------------
no_mutation_score=0 #round(overall_min_score)-2
score_type=CADD_raw_rankscore
##########################

wanted_patient_list=get_patients_list()

#Note: main_dict[sample_id][gene]=[cadd_score]
main_dict={}
all_genes=set()
overall_min_score=0
overall_max_score=0

vcf_files_path='/storage/gudalab/svural/TCGA_based_works/TCGA_BRCA_protected_d
ata/WUSM__Automated_Mutation_Calling/final_vcfs/'

for vcf_file in glob.glob(vcf_files_path+'*.vcf'):
```

```
  path,vcf_file_name = os.path.split(vcf_file)
  sample_id=vcf_file_name.split('.')[0].split('-')[2]

  if not sample_id in wanted_patient_list: continue

  vcf_file=open(vcf_file)
  skip_header_lines(vcf_file)
  #-------one sample_id------------
  for line in vcf_file:
    line=line.split()
    ########################################
    gene  =line[4].upper()
    score =float(line[score_type])
    ########################################

    #-----------------------------------
    if float(score) < overall_min_score:
      overall_min_score=score
    if float(score) > overall_max_score:
      overall_max_score=score
    #-----------------------------------

    all_genes.add(gene)

    if not sample_id in main_dict:            main_dict[sample_id]={}
    if not gene in main_dict[sample_id]:      main_dict[sample_id][gene]=[]

    main_dict[sample_id][gene].append(score)

stop_time=time.time()

print 'main_dict is loaded, ',
print 'took %.2f minutes' % ( (stop_time-start_time)/60.0 )
print 'number of genes:',len(all_genes)
print 'min value:',overall_min_score, 'no mutation score:',no_mutation_score
print 'max value:',overall_max_score
#*********************************************************************

if score_type==CADD_raw:
  score_offset=abs(overall_min_score)
  for sample in main_dict:
    for gene in main_dict[sample]:
      main_dict[sample][gene]=[score+score_offset for score in main_dict[sample][gene]]

#write output
```

```
#make header
header='Sample_id\t'
for gene in all_genes:
  header+=gene+'\t'
header=header[:-1]+'\n'

text=header
count=0
for sample in main_dict:
  count+=1
  print str(count)+' of '+str(len(main_dict))+' is done\r',
  line=sample+'\t'
  for gene in all_genes:
    if not gene in main_dict[sample]: # if gene does not have any mutations in this sample
      score=no_mutation_score
    else:
      score=sum(main_dict[sample][gene])
    line+=str(score)+'\t'
  line=line[:-1]+'\n'
  text+=line

open('raw.tsv','w').write(text)


stop_time=time.time()
print 'whole proceses took %.2f minutes' % ( (stop_time-start_time)/60.0 )
```

Running NMF algorithm

The following R script loads clinical information (stage) of the patients, runs some preprocessing steps and applies NMF algorithm. Here NMF algorithm is set to use 100 CPUs in parallel to run the algorithm for 100 iterations. Finally R script plots some essential NMF figures, saves patient-to-cluster assignments and clustering metrics, and exits.

```
Rscript used to run NMF algorithm
#!/usr/local/bin/Rscript --vanilla

suppressMessages(library(NMF))
```

```
##load phenotype data----------------------------------------
phenotype_file="/storage/gudalab/svural/TCGA_based_works/TCGA_BRCA_protected_
data/clinical/patient_stages.tsv"
phenotype_data=read.csv(phenotype_file,header=TRUE,sep="\t")
covariates <- data.frame( Stage=phenotype_data$stage )

#load mutation data-----------------------------------------
args<-commandArgs(TRUE)
my_file=args[1]
r=as.numeric(args[2]) ##number of clusters
data <- read.csv(my_file,header=TRUE,sep="\t")

#data preprocess-------------------------------------------
rownames(data) <- data[,1]
data <- data[,2:ncol(data)]
data <- t(data) ##because of naming difference
#---------------------------------------------------------

#run nmf
my_res <- nmf(data,r,nrun=100,seed=123456,.opt="v1p100")

#---------------Plot matrices---------------------------------
#Specify colors
Stage=c("red","green","black")
names(Stage)=c("Late Stage","Early Stage","Other")
ann_colors= list(Stage= Stage)

#plot consensus matrix
consensusmap(my_res,
          annCol=covariates,annColors=ann_colors,
          Rowv=F,Colv=F,
          tracks=NA,
          filename=paste(my_file,".consensus.jpg",sep=''),width=10,height=10
          )

#plot coefficient matrix
coefmap(my_res,
       Rowv=F,Colv=F,
       tracks=NA,
       filename=paste(my_file,".coefmap.jpg",sep=''),width=10,height=10
       )

#plot basis matrix
basismap(my_res,
        Rowv=F,Colv=F,
        tracks=NA,
```

```
        legend = FALSE,
        filename=paste(my_file,".basis.jpg",sep=''),width=10,height=10
        )
#------------------------------------------------------------

#write clusterring stats
my_summary<-summary(my_res)
summary_table <- rbind(dim(data)[1],as.data.frame(my_summary))
rownames(summary_table)[1] <- "num_features"
write.table(summary_table,paste(my_file,'.summary.tsv',sep=''),col.names=my_file)

cat("done\n")
```

Filter and Sort Data by Variance

The following Python script is used for feature selection by gene variance. The script accepts two command line input arguments. The first argument indicates the file name and the second argument is used to select the top variant genes (e.g. top 100 genes).

```
#!/usr/bin/python
import sys,os,argparse,time,glob,csv
import numpy
import functions
start_time=time.time()

load_data=functions.load_data

argument_parser = argparse.ArgumentParser()
argument_parser.add_argument('input_file')
argument_parser.add_argument('number_of_genes')

args = argument_parser.parse_args()

input_file=open(args.input_file)

data,patient_id_list,gene_list=load_data(input_file)

data_variance_values=data.var(axis=0).tolist()[0]
gene_variance_dict=dict(zip( gene_list , data_variance_values))
#gene_variance_dict[gene]=variance_value

data_column_sums=data.sum(axis=0).tolist()[0]
```

```
gene_column_sum_dict=dict(zip( gene_list , data_column_sums))

zero_percentage_dict={}
num_rows=float(numpy.size(data,axis=0))  #axis=0 -> y axis (num. of patients)    axis=1
-> x axis (num. of genes)
num_cols=numpy.size(data,axis=1)
for i in range(num_cols):
  gene_name=gene_list[i]
  zero_percentage_dict[gene_name]=(1-
(numpy.count_nonzero(data[:,i])/num_rows))*100

sorted_gene_list=sorted(gene_variance_dict.items(), key=lambda x: x[1], reverse=True)
#[ ('gene1',9.0),('gene2',8.1),('gene3',5)] sorted by variance value


##f_out=open('feature_variance_colSum_zeroPercent.tsv','w')
##f_out.write('Gene\tVariance\tCol.Sum\tZero_Percentage\n')
##for item in sorted_gene_list:
##  gene_name              =item[0]
##  gene_variance          =str(item[1])
##  column_sum             =str(gene_column_sum_dict[gene_name])
##  gene_zero_percentage =str(zero_percentage_dict[gene_name])
##
##  f_out.write(gene_name +'\t'+ gene_variance +'\t'+
column_sum+'\t'+gene_zero_percentage+'\n' )
##f_out.close()
##
##sys.exit()
###
#for number_of_genes in range(1,1001):
#print str(number_of_genes)+'\r',
#combine output data----------------------------------------------
number_of_genes=int(args.number_of_genes)
out_data=data[:,gene_list.index(sorted_gene_list[0][0])]
for i in range(1,number_of_genes):
  wanted_column_data=data[:,gene_list.index(sorted_gene_list[i][0])]
  out_data=numpy.append(out_data,wanted_column_data,axis=1)
#-----------------------------------------------------------------

#write output ----------------------------------------------------
header='PATIENT_ID\t'+'\t'.join([item[0] for item in
sorted_gene_list[:number_of_genes]])+'\n'
text=header
for row in range(len(patient_id_list)):
  patient_id=patient_id_list[row]
  data_line ='\t'.join( str(item) for item in out_data[row,:].tolist()[0])
```

```
  text+=patient_id+'\t'+data_line+'\n'
text=text[:-1]
open(str(number_of_genes)+'.cluster','w').write(text)
#------------------------------------------------------------------

stop_time=time.time()
print 'whole proceses took %.2f seconds' % (stop_time-start_time)
```

Order Data

This Python scripts orders the data according to various parameters.

```
#!/usr/bin/python
'''
takes a cluster formated data file orders samples by thier clusters and orders columns
(genes) by chromosomes and by variance/location in chromosome inside the
chromosomes
'''
import argparse,functions,os,sys,numpy

chr_gene_mapping_file='/storage/gudalab/svural/TCGA_based_works/TCGA_BRCA_pr
otected_data/WUSM__Automated_Mutation_Calling/chr_gene_mapping/uniq_sorted_all
_genes_with_pos'

argument_parser = argparse.ArgumentParser()
argument_parser.add_argument('input_file')  #cluster file, to be ordered
argument_parser.add_argument('sample_prediction_file')
args = argument_parser.parse_args()

input_file=open(args.input_file)
data,patient_id_list,gene_list=functions.load_data(input_file)

#============================================================
chr_gene_dict,ordered_all_genes_list=functions.read_gene_chromosome_mapping(chr_g
ene_mapping_file) #chr_gene_dict[chr1]=[gene1,gene2,gene3, ...]   genes are sorted by
chromosomal location
#------------------------------------------------------------ a trick to get a list of genes in the
chromosomal order
for gene in  ordered_all_genes_list:
  if gene not in gene_list:
    ordered_all_genes_list[ordered_all_genes_list.index(gene)]=0
ordered_gene_list=[i for i in ordered_all_genes_list if i!=0]
#------------------------------------------------------------------------------------------------------
-------------
```

```
#first order genes
out_data=data[:,gene_list.index(ordered_gene_list[0])]
for item in ordered_gene_list[1:]:
  wanted_column_data=data[:,gene_list.index(item)]
  out_data=numpy.append(out_data,wanted_column_data,axis=1)
#============================================================


#============================================================
sample_predictions_dict=functions.get_clusters_patients_list(
args.sample_prediction_file )  #d[cluster1]=[patient1, patient2, patient3 ...]
ordered_patient_list=[]
for i in sorted(sample_predictions_dict):
  ordered_patient_list+=sample_predictions_dict[i]

#then order rows/patients
out_data2=out_data[patient_id_list.index(ordered_patient_list[0]),:]
for item in ordered_patient_list[1:]:
  wanted_row_data=out_data[patient_id_list.index(item),:]
  out_data2=numpy.append(out_data2,wanted_row_data,axis=0)

sample_predictions_dict=functions.read_sample_predictions(
args.sample_prediction_file ) #d[cluster1]=[patient1, patient2, patient3 ...]

#============================================================


#write output ----------------------------------------------------
header='PATIENT_ID\t'+'\t'.join( ordered_gene_list )+'\t'+'PATIENT_CLUSTER'+'\n'
text=header
for row in range(len(patient_id_list)):
  patient_id=ordered_patient_list[row]
  data_line ='\t'.join( str(item) for item in out_data2[row,:].tolist()[0])
  patient_cluster=sample_predictions_dict[patient_id]
  text+=patient_id+'\t'+data_line+'\t'+patient_cluster+'\n'
text=text[:-1]

head,tail=os.path.split(args.input_file)
if head=='': head='.'
new_file_name=head+'/'+tail.split('.')[0]+'_ordered.cluster'
open(new_file_name,'w').write(text)
#-----------------------------------------------------------------

print 'Done'
```

REFERENCES

Adams, J. (2003). Potential for proteasome inhibition in the treatment of cancer. *Drug Discovery Today*, *8*(7), 307–315. http://doi.org/10.1016/S1359-6446(03)02647-3

Adeyinka, A., Nui, Y., Cherlet, T., & Snell, L. (2002). Activated mitogen-activated protein kinase expression during human breast tumorigenesis and breast cancer progression. *Clinical Cancer …*, *8*(June), 1747–1753. Retrieved from http://clincancerres.aacrjournals.org/content/8/6/1747.short

Adzhubei, I. a, Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–9. http://doi.org/10.1038/nmeth0410-248

Ali, H. R., Rueda, O. M., Chin, S.-F., Curtis, C., Dunning, M. J., Aparicio, S. A., & Caldas, C. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biology*, *15*(8), 431. http://doi.org/10.1186/s13059-014-0431-1

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., … Thomson, E. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, *474*(7353), 609–615. http://doi.org/10.1038/nature10166

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B …*, *57*(1), 289–300. Retrieved from http://www.jstor.org/stable/2346101

Breast cancer stages, retrieved from http://www.cancercenter.com/breast-cancer/stages/. (n.d.). Retrieved from http://www.cancercenter.com/breast-cancer/stages/

Breast Cancer Stages, retrieved from http://www.nationalbreastcancer.org/breast-cancer-stages. (n.d.). Retrieved from http://www.nationalbreastcancer.org/breast-cancer-stages

Breast Cancer Staging and Stages, retrieved from http://ww5.komen.org/BreastCancer/StagingofBreastCancer.html. (n.d.). Retrieved from http://ww5.komen.org/BreastCancer/StagingofBreastCancer.html

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://doi.org/10.1023/A:1010933404324

Bruna, A., Greenwood, W., Le Quesne, J., Teschendorff, A., Miranda-Saavedra, D., Rueda, O. M., … Caldas, C. (2012). TGFβinduces the formation of tumour-initiating cells in claudinlow breast cancer. *Nature Communications*. http://doi.org/10.1038/ncomms2039

Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, *101*(12), 4164–9. http://doi.org/10.1073/pnas.0308531101

Cardoso, F., Van't Veer, L., Rutgers, E., Loi, S., Mook, S., & Piccart-Gebhart, M. J. (2008). Clinical application of the 70-gene profile: the MINDACT trial. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *26*(5), 729–35. http://doi.org/10.1200/JCO.2007.14.3222

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., … Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States*

of America, *106*(45), 19096–19101. http://doi.org/10.1073/pnas.0910672106

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. http://doi.org/10.4161/fly.19695

Cooper, G. M., Stone, E. a, Asimenos, G., Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, *15*(7), 901–13. http://doi.org/10.1101/gr.3577405

Cornish, A., & Guda, C. (2015). A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *BioMed Research International*, *In press*, 456479. http://doi.org/10.1155/2015/456479

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., … Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, *486*(7403), 346–52. http://doi.org/10.1038/nature10983

Dawson, S. J., Provenzano, E., & Caldas, C. (2009). Triple negative breast cancers: clinical and prognostic implications. *European Journal of Cancer (Oxford, England : 1990)*, 27–40. http://doi.org/10.1016/S0959-8049(09)70013-9

DeLuca, A. (2013). Computational methods for efficient exome sequencing-based genetic testing. *Theses and Dissertations*. Retrieved from http://ir.uiowa.edu/etd/2473(Doctoral dissertation)

Elston, C., Ellis, I., & Pinder, S. (1999). Pathological prognostic factors in breast cancer.

Critical Reviews in Oncology/hematology, *31*, 209–223. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Pathological+prognostic+factors+in+breast+cancer#0

Forbes, S. a, Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., … Campbell, P. J. (2014). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, *43*(October 2014), 805–811. http://doi.org/10.1093/nar/gku1075

Furuuchi, K., Tada, M., Yamada, H., Kataoka, A., Furuuchi, N., Hamada, J., … Moriuchi, T. (2000). Somatic Mutations of the APC Gene in Primary Breast Cancers. *The American Journal of Pathology*, *156*(6), 1997–2005. http://doi.org/10.1016/S0002-9440(10)65072-9

Gaujoux, R., & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, *11*, 367. http://doi.org/10.1186/1471-2105-11-367

Güngör, C., Zander, H., Effenberger, K. E., Vashist, Y. K., Kalinina, T., Izbicki, J. R., … Bockhorn, M. (2011). Notch signaling activated by replication stress-induced expression of midkine drives epithelial-mesenchymal transition and chemoresistance in pancreatic cancer. *Cancer Research*, *71*(14), 5009–19. http://doi.org/10.1158/0008-5472.CAN-11-0036

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations*, *11*, 10–18. http://doi.org/10.1145/1656274.1656278

Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., … Bast, R. C. (2007). American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *25*(33), 5287–312. http://doi.org/10.1200/JCO.2007.14.2364

Hennessy, B., & Gonzalez-Angulo, A. (2009). Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. *Cancer Research*, *69*(10), 4116–4124. http://doi.org/10.1158/0008-5472.CAN-08-3441.Characterization

Hofree, M., Shen, J. P., Carter, H., Gross, A., & Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature Methods*, *10*(11), 1108–15. http://doi.org/10.1038/nmeth.2651

Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., … Perou, C. M. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, *7*, 96. http://doi.org/10.1186/1471-2164-7-96

International Agency for Research on Cancer. (2013). *Press Release No:223*.

International Agency for Research on Cancer. (2014). *The World Cancer Report 2014*. Retrieved from http://www.iarc.fr/en/publications/books/wcr/wcr-order.php

IPA; Ingenuity Systems Inc.; Redwood, CA, USA. (n.d.). Redwood, CA, USA: Ingenuity Systems Inc.

Kim, M. H., Seo, H. J., Joung, J.-G., & Kim, J. H. (2011). Comprehensive evaluation of matrix factorization methods for the analysis of DNA microarray gene expression

data. *BMC Bioinformatics*, *12 Suppl 1*, S8. http://doi.org/10.1186/1471-2105-12-S13-S8

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–5. http://doi.org/10.1038/ng.2892

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., … Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–76. http://doi.org/10.1101/gr.129684.111

Konecny, G., Pauletti, G., Pegram, M., Untch, M., Dandekar, S., Aguilar, Z., … Slamon, D. J. (2003). Quantitative Association Between HER-2/neu and Steroid Hormone Receptors in Hormone Receptor-Positive Primary Breast Cancer. *JNCI Journal of the National Cancer Institute*, *95*(2), 142–153. http://doi.org/10.1093/jnci/95.2.142

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–9. http://doi.org/10.1038/nmeth.1923

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., … Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics (Oxford, England)*, *28*(3), 311–7. http://doi.org/10.1093/bioinformatics/btr665

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–8.

http://doi.org/10.1038/nature12213

Lee, D., & Seung, H. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing …*, (1).

Lee, Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P., & Marth, G. T. (2014). MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. *PLoS ONE*, *9*(3), e90581. http://doi.org/10.1371/journal.pone.0090581

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*, *121*, 2750–2767. http://doi.org/10.1172/JCI45014

Lester, S., & Bose, S. (2009). Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Archives of Pathology & Laboratory Medicine*.

Li, C. I., Uribe, D. J., & Daling, J. R. (2005). Clinical characteristics of different histologic types of breast cancer. *British Journal of Cancer*, *93*(9), 1046–52. http://doi.org/10.1038/sj.bjc.6602787

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Preprint arXiv*, *00*(00), 3. http://doi.org/arXiv:1303.3997 [q-bio.GN]

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–60.

http://doi.org/10.1093/bioinformatics/btp324

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–8. http://doi.org/10.1101/gr.078212.108

Lim, E., Vaillant, F., Wu, D., Forrest, N. C., Pal, B., Hart, A. H., … Lindeman, G. J. (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nature Medicine*, *15*, 907–913. http://doi.org/10.1038/nm.2000

List, M., Hauschild, A.-C., Tan, Q., Kruse, T. a, Mollenhauer, J., Baumbach, J., & Batra, R. (2014). Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *Journal of Integrative Bioinformatics*, *11*(2), 236. http://doi.org/10.2390/biecoll-jib-2014-236

Liu, Y., Popp, B., & Schmidt, B. (2014). CUSHAW3: sensitive and accurate base-space and color-space short-read alignment with hybrid seeding. *PloS One*, *9*(1), e86869. http://doi.org/10.1371/journal.pone.0086869

Malhotra, G. K., Zhao, X., Band, H., & Band, V. (2010). Histological, molecular and functional subtypes of breast cancers. *Cancer Biology & Therapy*, *10*(10), 955–960. http://doi.org/10.4161/cbt.10.10.13879

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, *26*(16), 2069–2070. http://doi.org/10.1093/bioinformatics/btq330

Mitchell, T. M. (1997). *Machine Learning*. *Machine Learning* (Vol. 1). http://doi.org/10.1007/BF00116892

Ng, P., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research*, 863–874.

Paik, S., Shak, S., Tang, G., & Kim, C. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England Journal of Medicine*, 2817–2826.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., … Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *27*(8), 1160–7. http://doi.org/10.1200/JCO.2008.18.1370

Perou, C., Sørlie, T., & Eisen, M. (2000). Molecular portraits of human breast tumours. *Nature*, *533*, 747–752. Retrieved from http://www.nature.com/nature/journal/v406/n6797/abs/406747a0.html

Platt, J. (1998). Sequential minimal optimization: a fast algorithm for training support vector machines. 1998. *URL Citeseer. Ist. Psu. edu/platt98sequential. Html*, 1–21.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., … Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research : BCR*, *12*(5). http://doi.org/10.1186/bcr2635

RDevelopment, C. (2012). R: A Language and Environment for Statistical Computing. *Vienna, Austria*, *1*.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial …*, 41–46. Retrieved from http://www.researchgate.net/profile/Irina_Rish/publication/228845263_An_empirical_study_of_the_naive_Bayes_classifier/links/00b7d52dc3ccd8d692000000.pdf

Ritchie, G. R. S., & Flicek, P. (2014). Computational approaches to interpreting genomic sequence variation. *Genome Medicine*, *6*(10), 1–11. http://doi.org/10.1186/s13073-014-0087-1

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. http://doi.org/10.1016/0377-0427(87)90125-7

Salzberg, S. (1994). Book Review: C4. 5: Programs for machine learning by J. Ross Quinlan. Inc., 1993., (16), 235–240. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Book+Review+:+C4+.+5+:+Programs+for+Machine+Learning#2

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–7. http://doi.org/10.1073/pnas.74.12.5463

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, *7*(8), 575–576. http://doi.org/10.1038/nmeth0810-575

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids*

Research, *29*, 308–311. http://doi.org/10.1093/nar/29.1.308

Sørlie, T., & Perou, C. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, *98*(19). Retrieved from http://www.pnas.org/content/98/19/10869.short

Sørlie, T., & Tibshirani, R. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, *100*(14), 8418–8423. Retrieved from http://www.pnas.org/content/100/14/8418.short

Sparano, J. a, & Paik, S. (2008). Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *26*(5), 721–8. http://doi.org/10.1200/JCO.2007.15.1068

Stevens, K. N., Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor, *IT-13*(1), 21–27.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., & Ebert, B. L. (2005). Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles, *102*(43), 15545–15550. http://doi.org/10.1073/pnas.0506580102

TCGA. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70. http://doi.org/10.1038/nature11412

Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O., & Caldas, C. (2007). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biology*.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, *99*, 6567–6572. http://doi.org/10.1073/pnas.082099299

Veer, L. van't, Dai, H., & Vijver, M. Van De. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(345). Retrieved from http://www.nature.com/nature/journal/v415/n6871/abs/415530a.html

Vijver, M. Van De, & He, Y. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, *347*(25), 1999–2009.

Wagner, A. H. (2014). Computational methods for identification of disease-associated variations in exome sequencing. Retrieved from http://ir.uiowa.edu/cgi/viewcontent.cgi?article=5520&context=etd(Doctoral dissertation)

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164. http://doi.org/10.1093/nar/gkq603

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., … Zhao, Z. (2013). Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Medicine*, *5*(10), 91. http://doi.org/10.1186/gm495

Won Jeong, K., Chodankar, R., Purcell, D. J., Bittencourt, D., & Stallcup, M. R. (2012). Gene-specific patterns of coregulator requirements by estrogen receptor-α in breast cancer cells. *Molecular Endocrinology (Baltimore, Md.)*, *26*(6), 955–66.

http://doi.org/10.1210/me.2012-1066

Zheng, C. H., Zhang, L., Ng, V. T. Y., Shiu, S. C. K., & Huang, D. S. (2011). Molecular pattern discovery based on penalized matrix decomposition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*, 1592–1603. http://doi.org/10.1109/TCBB.2011.79

Zimmerman, E. (2014). Top of the 50 Smartest Company List: Illumina. *MIT Technology Review*. Retrieved from http://www.technologyreview.com/featuredstory/524531/why-illumina-is-no-1